第一章 预 篇

§1 引 言

数值代数是计算数学的重要分支，它是研究各类代数问题的数值解法的理论，例如矩阵的变换和分解、线性和非线性代数方程的解法、矩阵的特征值及特征向量的求解问题等.

量子力学、结构力学、网络分析、大地测量、机械振动、管理科学等应用技术学科中提出的大量的计算问题，大都直接归属于计算数学. 而计算数学各分支中的问题，如有限元方法、谱分解理论、稀疏矩阵问题、最小二乘法问题、数学规划与组合数学问题又都直接或间接成为数值代数的重要源泉和广阔的应用场所. 因此，数值代数的研究和发展不仅直接推动着计算数学的发展而且也是其它技术学科发展的重要前提.

古典代数理论，对于代数问题解的研究，已日臻完善；但当未知数的个数较多时，其数值解的获得，则仍是一个困难的问题. 即使在快速电子计算机问世以来，这个问题所面临的挑战，也并未因此而减弱. 相反，随着科学技术的发展，当生产斗争、科学实验日益迫切地要求人类向高维问题进军的时候，新的、富有成效的、收敛性能和稳定性能都更为良好的各种算法的研制，更是迫切地提到日程上来. 和其它计算数学分支一样，数值代数也随着近代物质文明的需要而日新月异地改变着自己的面貌.

本书作为数值分析的后继课程，将结合数值代数的近代发

展，把数值分析中有关数值代数的章节，适当地给于扩充和引申，同时也为一些新的实用的方法的学习和探索提供初步的导引。对于重要的方法，将以接近语句的形式把它们写出来。自然，算法描述是理论分析和推理演算的结晶，也是计算的有效性和经济性的统一，而不是规则和条文的堆砌。

本章将讨论一些在以后各章节中都要用到的工具和定理，它们包括：初等矩阵、矩阵的分解以及有关特征扰动方面的问题。

今将全书统一使用的符号一并在此列出。除维数和足码以外，通常用小写希腊字母代表数，用小写黑体拉丁字母代表向量，用大写的希腊或拉丁字母代表矩阵；A^T 表示矩阵 A 的转置，A^H 则表示它的共轭转置；用 $\boldsymbol{R}^n$、$\boldsymbol{C}^n$ 分别表示 n 维实的和复的线性空间，用 $\boldsymbol{R}^{n\times m}$、$\boldsymbol{C}^{n\times m}$ 分别表示 n 行 m 列的实或复的矩阵所成的空间；矩阵 A 的列一般记为 $\boldsymbol{a}_i$，而以 α_{ij} 或 a_{ij} 表示其在位置 (i,j) 的元素；用 $\boldsymbol{e}_i$ 表示单位矩阵 I 的第 i 列，而

$$\boldsymbol{e} = \Sigma \boldsymbol{e}_i$$

则表示分量全为 1 的向量；$|\boldsymbol{x}|$、$|A|$ 分别表示用向量 $\boldsymbol{x}$ 和矩阵 A 的分量的绝对值做成的向量和矩阵。在不致发生误解时，单维阵的维数、零阵或零向量的维数将不特别指明。用 $\det(A)$、$\lambda(A)$ 分别表示矩阵 A 的行列式和它的一个特征值，而用

$$\operatorname{tr}(A) = \sum_{i=1}^{n} \alpha_{ii} \qquad (1.1.1)$$

和

$$\rho(A) = \max|\lambda(A)| \qquad (1.1.2)$$

分别表示 A 的**迹**和**谱半径**；用

$$A\begin{pmatrix} i_1, & i_2, & \cdots, & i_p \\ j_1, & j_2, & \cdots, & j_p \end{pmatrix} \qquad (1.1.3)$$

表示选自 A 的第 $i_1, \cdots, i_p$ 各行和 $j_1, \cdots, j_p$ 各列的交点处的元素所成的行列式。

在不作特殊解释时，向量范数$\|\boldsymbol{x}\|$恒指 l_2 空间中的范数，即

$$\|\boldsymbol{x}\| = (\boldsymbol{x}^H\boldsymbol{x})^{1/2}, \tag{1.1.4}$$

而矩阵范数$\|A\|$是指与向量范数相容或从属于它的一种范数，即

$$\|A\boldsymbol{x}\| \leqslant \|A\|\|\boldsymbol{x}\| \tag{1.1.5}$$

或再加上一个限制：(1.1.5) 中的等号至少对一个 $\boldsymbol{x}\neq O$ 成立.

§2 初等矩阵

我们把形如

$$E(\boldsymbol{u}, \boldsymbol{v}, \sigma) = I - \sigma\boldsymbol{u}\boldsymbol{v}^H \tag{1.2.1}$$

的矩阵叫作**初等阵**，其中 $\boldsymbol{u}$、$\boldsymbol{v}$ 是 n- 维向量，σ 是数. 该阵是一个单位阵与一个秩 1 矩阵的组合. 当要把一个一般形式的矩阵约化为某种标准形式或任何压缩形式时，常需对矩阵进行一系列简单的相似变换，而这种工作通常是借助初等阵来实现的.

首先指出在 (1.2.1) 中，当视 σ 为参数，$\boldsymbol{u}$、$\boldsymbol{v}$ 为固定向量时，矩阵类 (1.2.1) 对乘法是封闭的. 实际上

$$E(\boldsymbol{u},\boldsymbol{v},\sigma)E(\boldsymbol{u},\boldsymbol{v},\tau) = E(\boldsymbol{u},\boldsymbol{v},\lambda), \tag{1.2.2}$$

其中，

$$\lambda = \sigma\tau\boldsymbol{v}^H\boldsymbol{u} - \sigma - \tau. \tag{1.2.3}$$

由于 $E(\boldsymbol{u}, \boldsymbol{v}, 0) = I$，令 $\lambda = 0$ 时解得

$$\tau = -\sigma(1-\sigma\boldsymbol{v}^H\boldsymbol{u}),\quad 1-\sigma\boldsymbol{v}^H\boldsymbol{u}\neq 0, \tag{1.2.4}$$

(1.2.4) 同时给出了 $E(\boldsymbol{u}, \boldsymbol{v}, \sigma)$ 有逆的条件和逆阵中的 τ 的计算.

直接计算$E(u, v, \sigma)$的行列式，还可进一步证明

$$\det E(u, v, \sigma) = 1 - \sigma v^H u. \qquad (1.2.5)$$

在消元中常见的有如下的初等阵。

2.1 行交换矩阵

$$P_{ij} = E(e_i - e_j, e_i - e_j, 1)$$

除第i行、j行、i列及j列上的4个元素有如下分布外，其余元素与单位矩阵I相同，称为**行交换矩阵**。

	i列	j列	
	0	1	i行
	1	0	j行

由于：$P_{ij}P_{ij} = I$及P_{ij}实对称，故P_{ij}是简单的正交阵。当用P_{ij}左乘某一矩阵时，其结果将使原矩阵的i、j两行互换，右乘时则将使i、j两列互换。

2.2 排列矩阵

$$P_{\alpha_1, \cdots, \alpha_n}$$

称为**排列矩阵**，其中$(\alpha_1, \cdots, \alpha_n)$代表$(1, 2, \cdots, n)$的一个排列，其元素，除位于

$$(i, \alpha_i), \quad i = 1, 2, \cdots, n$$

的元素规定为1外，其余全为零。该矩阵在每个行和列上都恰有一个1，因此，其转置阵也是排列阵，并且

$$P^{-1}_{\alpha_1, \cdots, \alpha_n} = P^T_{\alpha_1, \cdots, \alpha_n}, \qquad (1.2.6)$$

因而是正交阵，排列阵均可表示成P_{ij}型矩阵的乘积。它实际是单位阵经有限次换行的结果，因此其行列式的值均为±1。据定义，有

$$P_{\alpha_1, \cdots, \alpha_n} = \sum_{i=1}^{n} e_{\alpha_i} e_i^T. \qquad (1.2.7)$$

当用$P_{\alpha_1 \cdots \alpha_n}$左乘某矩阵时，其结果将把原矩阵的第$\alpha_i$行换到

第 i 行，右乘时，则结果是把原矩阵的第 α_i 列换到第 i 列。

2.3 消元矩阵

形如

$$L_j = E(\boldsymbol{l}_j, \boldsymbol{e}_j, 1), \qquad (1.2.8)$$
$$\boldsymbol{e}_k^T \boldsymbol{l}_j = 0, \; k \leqslant j$$

的矩阵，其中 $\boldsymbol{l}_j$ 是前 j 个分量为零的向量，设

$$\boldsymbol{l}_j^T = (0, \cdots, 0, l_{j+1,j}, \cdots, l_{n,j}),$$

则

$$L_j = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -l_{j+1,j} & 1 & & \\ & & \vdots & & \ddots & \\ & & -l_{n,j} & & & 1 \end{pmatrix} \qquad (1.2.9)$$

不难验证

$$L_j^{-1} = E(-\boldsymbol{l}_j, \boldsymbol{e}_j, 1), \; \det(L_j) = 1. \qquad (1.2.10)$$

当用 L_j 左乘于 $\boldsymbol{A}$ 时，其结果将把原矩阵中的第 j 行分别乘以 $-l_{j+1,j}, \cdots, -l_{n,j}$ 并依次加到矩阵的第 $j+1, \cdots, n$ 各行中，右乘时，$\boldsymbol{A}$ 的第 $j+1$ 至第 n 列将分别乘以 $-l_{j+1,j}, -l_{j+2,j}, \cdots, -l_{n,j}$ 并加到第 j 列，即左乘时将改变原矩阵的第 i 行以下的所有的行，右乘时只改变原矩阵的第 i 列。这种矩阵在求逆时，仅需把 $\boldsymbol{l}_i$ 换为 $-\boldsymbol{l}_i$，实际并不进行算术运算，因而并不引进舍入误差。不仅如此，在计算形如

$$L_1 L_2 \cdots L_k, \; L_1^{-1} L_2^{-1} \cdots L_k^{-1}$$

的矩阵乘积时，也不会引进误差。例如乘积矩阵 $L_1 \cdots L_K$ 的第 j 列只不过是 L_j 的第 j 列的重抄而已。类似地，若记

$$\boldsymbol{R}_i = L_i^T = E(\boldsymbol{e}_i, \boldsymbol{l}_i, 1),$$

可知

$$R_k^{-1}R_{k-1}^{-1}\cdots R_1^{-1}=\begin{pmatrix}1 & x & \cdots & x\\ & 1 & x\cdots & x\\ & & \ddots & \\ & & 1\cdots & x\\ 0 & & \ddots & 1\end{pmatrix}$$

应是单位上三角阵，其第 1 至 K 行将依次是 $l_1^T\cdots l_k^T$ 的重抄.

2.4 初等 Hermite 矩阵

矩阵 $$H=E(u,u,2),\quad \|u\|=1 \tag{1.2.11}$$

称为初等 Hermite 矩阵，文献上由于强调的侧面不同，有时又被称为 Householder 矩阵或镜像反射矩阵.

不难验证它既是Hermite 阵又是酉阵. 由于它还是一个对合阵，因而它的逆就是其自身. 据(1.2.5)及(1.2.11) 易知，$\det(H)=-1$.

由于下面三个定理，使这一类矩阵在矩阵的变换与分解中有着十分重要的应用.

定理1.2.1 对任意 $x\in C^n$，存在初等 Hermite 阵 H，使

$$Hx=\alpha e_1, \tag{1.2.12}$$

其中$|\alpha|=\|x\|$；特别，可取

$$\alpha=-\mathrm{sign}(\mathrm{Re}(\xi_1))\|x\|, \tag{1.2.13}$$

其中 ξ_1 是 $x^T=(\xi_1,\cdots,\xi_n)$ 的第 1 个分量.

证明 记 $H=I-2uu^H$，由

$$Hx=x-2(u^Hx)u=\alpha e_1,$$

故

$$u=\frac{(x-\alpha e_1)}{\|x-\alpha e_1\|}e^{i\theta}\ (\theta\text{ 是实数}),$$

但 uu^H 与 θ 无关，取 $\theta=0$. 考查

$$\|x-\alpha e_1\|^2=2(\|x\|^2-\mathrm{Re}(\alpha\xi_1)). \tag{1.2.14}$$

当按 (1.2.13) 确定 α 时，(1.2.14) 右端中方能有更多的有效数字.

作为本定理的应用，可证明下述著名的 **Schur** 定理.

定理1.2.2 任意矩阵可以经由酉相似变换化为上三角阵.

推论1.2.1 正规矩阵可以经由酉相似变换化为对角阵.

由于定理 1.2.2 及其推论的证明，并借助矩阵的特征值及特征向量，可知，对计算矩阵的特征问题来说，定理本身并非是构造性的，但它却是构造逐次逼近法求解矩阵特征问题的基础.

定理 1.2.2 及其推论的证明，已收入本章的习题. 在第三章中，我们就实矩阵的情形，给出了一个类似的更为适用的结果.

§3 矩阵的分解

首先，我们指出：对于任意的矩阵 $\boldsymbol{A}\in \boldsymbol{C}^{n\times m}$，设其秩为 p，则可以找到秩为 p 的两个矩阵 $\boldsymbol{M}$ 和 $\boldsymbol{P}$，其中 $\boldsymbol{M}$ 有 p 个列，$\boldsymbol{P}$ 有 p 个行，即 $\boldsymbol{M}\in \boldsymbol{C}^{n\times p}$ 而 $\boldsymbol{P}\in \boldsymbol{C}^{p\times m}$，使

$$\boldsymbol{A}=\boldsymbol{MP}. \tag{1.3.1}$$

我们将用两种方法说明上述论断.

3.1 列主元法

$$设\ \boldsymbol{A}=\boldsymbol{A}^{(0)}=(\boldsymbol{\alpha}_1^{(0)},\ \boldsymbol{\alpha}_2^{(0)},\ \cdots,\ \boldsymbol{\alpha}_m^{(0)})\neq 0. \tag{1.3.2}$$

令 j_1 为使 $\boldsymbol{\alpha}_j^{(0)}\neq 0$ 的最小的足码，且令

$$|\boldsymbol{e}_{i_1}^T\boldsymbol{\alpha}_{j_1}^{(0)}|=\max_i|\boldsymbol{e}_i^T\boldsymbol{\alpha}_{j_1}^{(0)}|, \tag{1.3.3}$$

即 $\boldsymbol{e}_{i_1}^T\boldsymbol{\alpha}_{j_1}^{(0)}$ 是 $\boldsymbol{\alpha}_{j_1}^{(0)}$ 的依模最大的分量，作

$$\tilde{A}^{(1)}+\boldsymbol{P}_{1i_1}\boldsymbol{A}^{(0)}=(\tilde{\boldsymbol{\alpha}}_1^{(0)},\ \tilde{\boldsymbol{\alpha}}_2^{(0)},\ \cdots,\ \tilde{\boldsymbol{\alpha}}_m^{(0)}), \tag{1.3.4}$$

记

$$\boldsymbol{l}_1^T=(1,\ l_{21},\ \cdots,\ l_{n1}),$$

$$l_{k1}=\boldsymbol{e}_k^T\tilde{\boldsymbol{\alpha}}_{j_1}^{(0)}/\boldsymbol{e}_1^T\tilde{\boldsymbol{\alpha}}_{j_1}^{(0)},\ k=2,\ \cdots,\ n, \tag{1.3.5}$$

$$L_1 = E(l_1, e_1, 1),$$
$$A^{(1)} = L_1 \tilde{A}^{(1)};$$

则在

$$A^{(1)} = (a_1^{(1)}, a_2^{(1)}, \cdots, a_m^{(1)}) \tag{1.3.6}$$

中，其第一行因 $e_1^T a_{j_1}^{(1)} \neq 0$ 必为非零行，但其前 j_1-1列上的元素以及 $a_{j_1}^{(1)}$ 的第一行以下的元素全为0。

如果 $A^{(1)}$ 的第一行以下全为零，则由

$$A^{(0)} P_{1j_1} L_1^{-1} A^{(1)} = P_{1j_1} L_1^{-1} A^{(1)} \tag{1.3.7}$$

及 $A_1 = e_1 e_1^T A_1$，有

$$L_1^{-1} A^{(1)} = (L_1^{-1} e_1)(e_1^T A^{(1)}).$$

可见 $A^{(0)}$已经表示成 1 个非零的列向量 $P_{1j_1} L_1^{-1} e_1$ 与一个非零的行向量 $e_1^T A^{(1)}$ 的乘积，此时 $A^{(0)}$ 的秩也只能是1。即所述论断是正确的。

如果 $A^{(1)}$ 的第 1 行以下第 j_1 列以右的矩阵 $\tilde{A}^{(2)}$ 非零，在 $A^{(0)}$的秩为 p 的前提下可知 $\tilde{A}^{(2)}$ 的秩应为$p-1$。对行 n 使用归纳法，可设

$$A^{(1)} = (A_1^{(1)} \vdots A_2^{(1)}) = \begin{pmatrix} 0 \cdots a_{1j_1}^{(1)} & \vdots & a_{1j_1+1}^{(1)} \cdots a_{1m}^{(1)} \\ \cdots\cdots\cdots & \vdots & \cdots\cdots\cdots\cdots \\ 0 & \vdots & \tilde{A}_2 \end{pmatrix}$$
$$= \begin{pmatrix} A_{1,1}^{(1)} & \vdots & A_{1,2}^{(1)} \\ \cdots\cdots & \cdots & \cdots\cdots \\ 0 & \vdots & \tilde{A}_2 \end{pmatrix},$$

而据归纳法可设$\tilde{A}_2 = \tilde{M}\tilde{P}$，这里 $\tilde{M}$，$\tilde{P}$ 都是秩为 $p-1$ 且各有 $p-1$ 列和 $p-1$ 行的阵，记 $\bar{P} = (0 \vdots \tilde{P})$，$\bar{P}$ 是秩为 $p-1$ 但有 $n-1$ 行和m列的阵。可得

$$A^{(1)} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} (A_{1,1}^{(1)} \cdots A_{1,2}) + \begin{pmatrix} 0 \\ \\ \\ \tilde{M} \end{pmatrix} \bar{P}$$

$$=\left(\begin{array}{c:c:c:c} 1 & 0 & A_{1,1}^{(1)} & A_{1,2}^{(1)} \\ 0 & & & \\ & & \cdots\cdots & \cdots\cdots \\ \vdots & \widetilde{M} & & \\ 0 & & 0 & \widetilde{P} \end{array}\right),$$

右端已约化为所需要的形式。由于

$$A^{(0)}=P_{1i_1}L_1^{-1}\begin{pmatrix} 1 & 0 & & \\ 0 & & & \\ \vdots & \widetilde{M} & & \\ 0 & & & \end{pmatrix}\begin{pmatrix} A_{1,1}^{(1)} & A_{1,2}^{(1)} \\ 0 & \widetilde{P} \end{pmatrix},$$

令

$$P_{1i_1}L_1^{-1}\begin{pmatrix} 1 & 0 \\ 0 & \\ \vdots & \widetilde{M} \\ 0 & \end{pmatrix}=M,\quad \begin{bmatrix} A_{1,1}^{(1)} & A_{1,2}^{(1)} \\ 0 & \widetilde{P} \end{bmatrix}=\mathbf{P},$$

可知

$$A^{(0)}=MP \tag{1.3.8}$$

具有所要求的形式。

在 $n=m$、$A^{(0)}$ 非奇异的前提下，如果进一步假定 $M=L$ 是单位下三角阵，$P=R$ 是上三角阵，此时，称 $A=A^{(0)}$ 有三角分解：

$$A=LR. \tag{1.3.9}$$

可证此种分解是唯一的。

3.2 初等 Hermite 阵法

这里，我们主要有以下结论。

设 $A\in\mathbf{C}^{n\times m}$ 且秩为 p，则存在秩为 p 的两个矩阵 W 和 R，它们各有 p 个列和 p 个行，其中 W 的列还是互相直交的单位向量，使

$$A=WR, \tag{1.3.10}$$

记

$$A = A^{(0)} = (\boldsymbol{a}_1^{(0)}, \cdots, \boldsymbol{a}_m^{(0)}), \tag{1.3.11}$$

令 $\boldsymbol{a}_{j_1}^{(0)}$ 是 $A^{(0)}$ 中的第一个非零列. 依定理1.2.1作初等Hermite阵

$$H_1 = I - 2\boldsymbol{u}_1\boldsymbol{u}_1^H \tag{1.3.12}$$

使

$$H_1\boldsymbol{a}_{j_1}^{(0)} = \gamma_{1j_1}\boldsymbol{e}_1, \quad |\gamma_{1j_1}| = \|\boldsymbol{a}_{j_1}^{(0)}\| \neq 0, \tag{1.3.13}$$

令

$$A^{(1)} = H_1A^{(0)}, \tag{1.3.14}$$

此时,$A^{(1)}$的前 j_1-1 列是零向量, 第 j_1 列除第 1 个分量 $\gamma_{1j_1}\neq 0$ 外，其余分量都是零。

如果$A^{(1)}$ 的第 1 行以下全为零，这时，由

$$A^{(0)} = H_1^{-1}A^{(1)} = H_1A^{(1)} = (H_1\boldsymbol{e}_1)(\boldsymbol{e}_1^TA^{(1)}) \tag{1.3.15}$$

以及 $H_1\boldsymbol{e}_1$ 是非零列向量，$\boldsymbol{e}_1^TA^{(1)}$是非零行向量和 $A^{(0)}$ 的秩 是 1，故已合于预期的结论。

如果 $A^{(1)}$ 的第 1 行以下第 j_1 列以右的矩阵 $\tilde{A}^{(2)}$ 非 零，用归纳法，可设

$$\tilde{A}^{(2)} = \tilde{H}^{(2)}\tilde{R}^{(2)}. \tag{1.3.16}$$

在$A^{(0)}$ 的秩为 p 的前提下,$\tilde{A}^{(2)}$、$\tilde{H}^{(2)}$、$\tilde{R}^{(2)}$的秩均应是$p-1$，且 $\tilde{H}^{(2)}$由 $p-1$ 个互相直交的单位向量组成，$\tilde{R}^{(2)}$由 $p-1$ 个线性无关的行向量组成. 这时，由

$$A^{(1)} = \begin{pmatrix} 0\cdots\ a_{1j_1}^{(1)} & \vdots\ a_{1,j_1+1}^{(1)}\ \cdots\ a_{1m}^{(1)} \\ \cdots\cdots & \cdots\cdots \\ 0 & \vdots\ \tilde{H}^{(2)}\tilde{R}^{(2)} \end{pmatrix} = \begin{pmatrix} A_{1,1}^{(1)} & \vdots\ A_{1,2}^{(1)} \\ \cdots\cdots & \cdots\cdots \\ 0 & \vdots\ \tilde{H}^{(2)}\tilde{R}^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} A_{1,1}^{(1)} \quad A_{1,2}^{(1)} \\ \cdots\cdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \cdots\cdots \\ \tilde{H}^{(2)}\tilde{R}^{(2)} \end{pmatrix}.$$

这里

$$\bar{R}^{(2)} = (\overbrace{0}^{j_1} \ \vdots\ \tilde{R}^{(2)})\}(p-1),$$

故

$$A^{(1)}=\begin{pmatrix} 1 & 0 \\ \cdots & \cdots \\ 0 & \\ \vdots & \widetilde{H}^{(2)} \\ 0 & \end{pmatrix}\begin{pmatrix} A_{1,1}^{(1)} & A_{1,2}^{(1)} \\ & \\ & \\ 0 & \widetilde{R}^{(2)} \end{pmatrix},$$

令

$$W=H_1\begin{pmatrix} 1 & 0 \\ \cdots & \cdots \\ 0 & \widetilde{H}^{(2)} \end{pmatrix},\quad R=\begin{pmatrix} A_{1,1}^{(1)} & A_{1,2}^{(1)} \\ 0 & \widetilde{R}^{(2)} \end{pmatrix},$$

则由 $A^{(0)}=H_1A^{(1)}$，有

$$A^{(0)}=A=WR \tag{1.3.17}$$

具有所要求的形式。

当 $n=m$，且 $A^{(0)}$ 非奇异时，如果进一步限制 R 为对角元为正数的上三角阵 U，则分解是存在且唯一的。

首先，当 A 非奇异时，由于 $A=WR$ 中，W、R 的秩均是 n，又 W 的各列是互相正交的单位向量，此时 W 是酉阵，R 可以写成三角阵的形式，它的对角元一定可以表示成正数，因为 R 的对角元为

$$\gamma_{kk}=|\gamma_{kk}|e^{i\theta_k},\ \theta_k\ \text{是实数} \tag{1.3.18}$$
$$k=1,2,\cdots,n$$

由

$$WR=(WD)(D^{-1}R)=QU \tag{1.3.19}$$

这里，$D=\text{diag}(e^{i\theta_1},\cdots,e^{i\theta_n})$，记 $D^{-1}R=U$ 是对角元为正数的上三角阵，又 D 是对角酉阵，于是 $Q=WD^{-1}$ 仍是酉阵。从而

$$A=QU, \tag{1.3.20}$$

其中 Q 是酉阵，U 是对角元为正数的上三角阵。分解的唯一性，则可由下面的分析说明，设

$$Q_1U_1=Q_2U_2,$$

则

$$Q_2^H Q_1 = U_2 U_1^{-1}$$

等式两端既是酉阵又是单位上三角阵，推知两端应同时是单位阵，于是 $Q_1 = Q_2$，$U_1 = U_2$。

3.3 三角分解的表现

当 $A \in C^{n\times n}$ 非奇异时，有可能得到下列的分解

$$A = LR,$$

其中 L 是单位下三角阵，R 是对角元非零的上三角阵，如果把 R 的对角元从 R 的左边提出来，可得

$$A = LDU, \tag{1.3.21}$$

其中 L 是单位下三角阵，U 是单位上三角阵，D 是非奇异对角阵。下面讨论怎样用 A 的元素来表示 L、D、U。

利用矩阵的分块表示法，设

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} D_{11} & \\ & D_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}, \tag{1.3.22}$$

这里，再设 A_{11}，L_{11}，U_{11} 都是 $p(\leqslant n)$ 阶方阵，有

$$A_{11} = L_{11} D_{11} U_{11},$$

$$\det(A_{11}) = \det(D_{11}) = \delta_1 \delta_2, \cdots, \delta_p, \tag{1.3.23}$$

$$D_{11} = \mathrm{diag}(\delta_1, \cdots, \delta_p),$$

$$\delta_p = A\begin{pmatrix} 1, 2, \cdots, p \\ 1, 2, \cdots, p \end{pmatrix} \Big/ A\begin{pmatrix} 1, 2, \cdots, p-1 \\ 1, 2, \cdots, p-1 \end{pmatrix}, \tag{1.3.24}$$

由

$$(A_{11},\ A_{12}) = L_{11}(D_{11}U_{11},\ D_{11}U_{12}); \tag{1.3.25}$$

由于两端都是行数为 p 的矩阵，故当两端取列号相同的 p 个列时，得出的行列式也应相等。特别，当两端取 1～$(p-1)$ 列并附加以第 j 列 $(j>p)$ 时，得

$$A\begin{bmatrix}1,2,\cdots,p-1,p\\1,2,\cdots,p-1,j\end{bmatrix}=\det(L_{11})\det(D_{11})U\begin{bmatrix}1,2,\cdots,p-1,p\\1,2,\cdots,p-1,j\end{bmatrix},$$

$$u_{p,j}=A\begin{bmatrix}1,2,\cdots,p-1,p\\1,2,\cdots,p-1,j\end{bmatrix}\Big/A\begin{bmatrix}1,2,\cdots,p\\1,2,\cdots,p\end{bmatrix}; \qquad (1.3.26)$$

这里，$u_{p,j}$ 是U的位于（p，j）的元素。比较

$$\begin{bmatrix}A_{11}\\A_{21}\end{bmatrix}=\begin{bmatrix}L_{11}&D_{11}&U_{11}\\L_{21}&D_{22}&U_{11}\end{bmatrix}$$

两端各有 n 个行和 p 个列，两端取前 $p-1$ 行和第 i 行形成行列式，由于 $\det(U_{11})=1$，得

$$l_{i,j}=A\begin{pmatrix}1,2\cdots,p-1,i\\1,2,\cdots,p-1,p\end{pmatrix}\Big/A\begin{pmatrix}1,\cdots,p\\1,\cdots,p\end{pmatrix}\ (i>p). \qquad (1.3.27)$$

§4 矩阵的 Jordan 标准型

设 $A\in C^{n\times n}$，据线性代数知识，存在非奇异矩阵H，使

$$H^{-1}AH=J=\begin{pmatrix}J_1&&&\\&J_2&&\\&&\ddots&\\&&&J_r\end{pmatrix}; \qquad (1.4.1)$$

其中

$$J_i=\begin{pmatrix}\lambda_i&1&&\\&\lambda_i&1&\\&&\ddots&\ddots\\&&&\lambda_i\end{pmatrix}, \qquad (1.4.2)$$

$\lambda_i(i=1,\cdots,r)$ 是A的特征值，J_i 为对应 λ_i 的上 Jordan 块，λ_i可能相异或相等。为确定起见，我们不妨规定：J 中对应同一特征值的 Jordan 块是彼此相邻的。于是，当用$J_s^{(k)}$的上标 k 表示A的相异特征值的编号，下标 s 表示对应同一特征值的 Jordan 块的编号时，则（1.4.1）可重写为

$$H^{-1}AH = \mathrm{diag}\,(J_1^{(1)}, \cdots, J_{i_1}^{(1)}, \cdots, J_1^{(r)}, \cdots, J_{i_r}^{(r)}). \quad (1.4.3)$$

我们把（1.4.1）和（1.4.3）统称为A的 **Jordan 标准型**，而称H为A的一个**变换矩阵**。

记$J_s^{(k)}$的阶数为$n_s^{(k)}$，则

$$n = \sum_{k=1}^{r} \sum_{s=1}^{i_k} n_s^{(k)}. \quad (1.4.4)$$

当每个$n_s^{(k)}$均为 1 时，A将相似于一个对角阵，即

$$H^{-1}AH = \mathrm{diag}\,(\lambda_1, \cdots, \lambda_n). \quad (1.4.5)$$

据习题（1.12），任意正规阵均酉相似于一对角阵，于(1.4.5)两端各用U^H和U去左乘和右乘，这里U是酉阵，则于右端得到正规阵，因此，凡相似于对角阵的阵，必也相似于一正规阵，所以，相似于对角阵的阵又被称为可正规化矩阵，但酉相似于对角阵的阵则必是正规阵。

一般情况下，由（1.4.1）有

$$AH = HJ = H\mathrm{diag}\,(J_1, \cdots, J_r);$$

记p为$J_s\,(1 \leqslant s \leqslant r)$的第 1 列在$J$中的编号，则由

$$AHe_p = Ah_p = \lambda_s h_p$$

这里h_p是H的第p列。上述表明，它是A的对应于λ_s的特征向量。当J_s的阶数大于 1 时，H的对应于J_s的其余的列，例如对应J_s的第 2，第3…列的列，由于满足

$$Ah_{p+1} = \lambda_s h_{p+1} + h_p,\ Ah_{p+2} = \lambda_s h_{p+2} + h_{p+1}, \cdots \quad (1.4.7)$$

因而都不是A的特征向量。据此，可知A的线性无关的特征向量的个数应为A的 Jordan 标准型中的 Jordan 块的个数，即r。于是，一般地有

$$r \leqslant n. \quad (1.4.8)$$

定义1.4.1 当$r < n$，即当矩阵A的线性无关的特征向量的个数小于矩阵的阶数时，称A为**亏损（defective）矩阵**；

反之，则称为非亏损的。

据此可知：可正规化的矩阵应是非亏损的。特别，当A的所有特征值均相异时，则A应是非亏损的。相似于正规阵、可正规化、非亏损这是不同说法的同一矩阵类。因其具有重要的应用，是需要着重进行研究的。

当A是非亏损矩阵时，由

$$H^{-1}AH=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$$

两端同用D^{-1}及D去左乘和右乘，这里D是非奇异的任意对角阵，可得

$$(HD)^{-1}A(HD)=\mathrm{diag}(\lambda_1,\cdots,\lambda_n), \qquad (1.4.9)$$

适当选择D可使HD的各列成为单位长向量，它们都是A的特征向量，其个数为n，且彼此线性无关，此时称A是具有完全标准化特征向量系的矩阵。如果仍把HD记为H，因此，当A非亏损时，我们总可以假设，它具有变换矩阵H，其各列就是A的一组完全标准的特征向量组成的，记

$$H=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n),\ \|\boldsymbol{x}_k\|=1, \quad (1\leqslant k\leqslant n). \qquad (1.4.10)$$

当H确定之后，H^{-1}也随之而确定，记

$$(H^{-1})^T=(\boldsymbol{g}_1,\ \boldsymbol{g}_2,\cdots,\boldsymbol{g}_n), \qquad (1.4.11)$$

$$\frac{\boldsymbol{y}_k}{s_k}=\boldsymbol{g}_k,\ \ |s_k|=1/\|\boldsymbol{g}_k\|,\ \ k=1,2,\cdots,n;$$

则$\boldsymbol{g}_k=(H^{-1})^T\boldsymbol{e}_k$，$\boldsymbol{g}_k^T=\boldsymbol{e}_k^T(H^{-1})$，由

$$H^{-1}A=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)H^{-1}$$

可得

$$\boldsymbol{g}_k^TA=\lambda_k\boldsymbol{g}_k^T,\ \ k=1,2,\cdots,n$$

称$\boldsymbol{g}_k$为A的对应于特征值λ_k的左特征向量，而$\boldsymbol{y}_k=\boldsymbol{g}_k/\|\boldsymbol{g}_k\|$为$A$的对应于特征值$\lambda_k$的标准化的左特征向量。由前面的分析可知，当$A$是非亏损的矩阵时，则存在$A$的右的和左的完全标准特征向量系。由于$H^{-1}H=I$，故

$$e_i^T H^{-1} H e_j = g_i^T x_j = \frac{1}{s_i} y_i^T x_j = \begin{cases} 0, & i \neq j; \\ 1, & i = j \end{cases}$$

或

$$y_i^T x_j = \begin{cases} 0, & i \neq j; \\ s_i \neq 0, & i = j, \end{cases} \tag{1.4.12}$$

即非亏损阵的左右完全标准特征向量系形成一个双正交系。

在（1.4.12）中，因 x_i，y_i 都是单位向量，据Cauchy 不等式，有

$$|s_i| \leqslant 1, \qquad i = 1, 2, \cdots, n. \tag{1.4.13}$$

当A是正规阵时，A的变换矩阵H可取为酉阵，此时，据(1.4.11）及 $H^{-1} = H^H$，知有 $g_i = \bar{x}_i (i = 1, \cdots, n)$，从而，

$$|s_i| = |g_i| = 1 \tag{1.4.14}$$

即正规阵是具有完全标准化的双正交特征向量系的矩阵。

注意，当A的特征值均相异时，则A显然相似于以A的特征值为对角元的对角阵，因而必为非亏损阵，它的对应于某一特征值的特征向量，如不计特征向量各分量间的比例因子，应是唯一的，其变换矩阵H，如不计列的次序，则也是唯一的。但当A是特征值均不相异的非亏损阵时，其变换矩阵H将不是唯一的，即A可以有不同的变换矩阵H能同时使A成为对角阵。例如，当A是单位阵I时，实际上任意的非奇异矩阵都可作为A的变换矩阵。

反之，如果同一的变换矩阵H可使两个不同的矩阵A、B同时化为对角阵，自然地应该问A、B之间存在什么关系，此时，因

$$\begin{aligned} H^{-1}AH &= \mathrm{diag}(\lambda_1, \cdots, \lambda_n), \\ H^{-1}BH &= \mathrm{diag}(\mu_1, \cdots, \mu_n) \end{aligned} \tag{1.4.15}$$

将可推出A、B可换，即

$$AB = BA, \tag{1.4.16}$$

进一步，可得下列定理。

定理1.4.1 非亏损矩阵A、B可换的充要条件是，A、B有相同的变换矩阵。

证明 只证条件的必要性。

由于A非亏损，故存在非奇异阵H，使

$$H^{-1}AH=\mathrm{diag}(\lambda_1,\cdots,\lambda_n),\quad AH=H\mathrm{diag}(\lambda_1,\cdots,\lambda_n),$$

记

$$\begin{aligned}&\mathrm{diag}(\lambda_1,\cdots,\lambda_n)=\mathrm{diag}(\lambda_1 I_{r_1},\lambda_2 I_{r_2},\cdots,\lambda_s I_{r_s}),\\&H=(\boldsymbol{h}_1^{(1)},\cdots,\boldsymbol{h}_{r_1}^{(1)},\boldsymbol{h}_1^{(2)},\cdots,\boldsymbol{h}_{r_2}^{(2)},\cdots,\boldsymbol{h}_1^{(s)},\cdots,\boldsymbol{h}_{r_s}^{(s)}),\end{aligned}\qquad(1.4.17)$$

即设A中的相等的特征值都被排列成相邻的，(1.4.17)中的诸$\lambda_1,\cdots,\lambda_s$视为相异的，而$\boldsymbol{h}_j^{(i)}$则表示$A$的对应于第$i$个相异特征值的第$j$个线性无关的特征向量。由于

$$ABH=BH\mathrm{diag}(\lambda_1 I_{r_1},\cdots,\lambda_s I_{r_s})$$

可见当$\boldsymbol{h}$是A的特征向量时，$B\boldsymbol{h}$也是A的特征向量，可知：$B\boldsymbol{h}_k^{(i)}$将属于$\boldsymbol{h}_1^{(i)},\cdots,\boldsymbol{h}_{r_i}^{(i)}$张成的空间，有

$$B\boldsymbol{h}_k^{(i)}=\sum_{j=1}^{r_i}\alpha_{jk}^{(i)}\boldsymbol{h}_j^{(i)},\qquad(1.4.18)$$

$$i=1,\cdots,s,\quad k=1,\cdots,r_i,$$

$$\therefore\qquad BH=H\mathrm{diag}(X_1,\cdots,X_s);\qquad(1.4.19)$$

其中$X_i(\alpha_{jk}^{(i)})$是(1.4.18)右端中的系数阵。由于B非亏损，故$H^{-1}BH=\mathrm{diag}(X_1,\cdots,X_s)$应能化为对角阵，即存在非奇异阵$P_i(i=1,\cdots,s)$使

$$P_iX_iP_i^{-1}=\Lambda_i\quad(i=1,\cdots,s);$$

其中Λ_i为对角阵，阶数为r_i，$(i=1,\cdots,s)$，令

$$P=\mathrm{diag}(P_1,\cdots,P_s),$$

则$(HP)^{-1}B(HP)=\mathrm{diag}(\Lambda_1,\cdots,\Lambda_s)$，又

$$\begin{aligned}(HP)^{-1}AHP&=P^{-1}\mathrm{diag}(\lambda_1 I_{r_1},\cdots,\lambda_s I_{r_s})P\\&=\mathrm{diag}(\lambda_1 I_{r_1},\cdots,\lambda_s I_{r_s}).\end{aligned}$$

可见A、B有相同的变换矩阵HP。

当A是亏损阵时，据（1.4.7）有定义：

定义1.4.2 设λ是A的特征值，$\boldsymbol{x}$是非零向量，l为某一正整数，如果

$$(A-\lambda I)^k\boldsymbol{x}\neq 0,\ k=1,\cdots,l-1;$$

但$(A-\lambda I)^l\boldsymbol{x}=0$，则称$\boldsymbol{x}$为$A$的对应于特征值$\lambda$的$l$**级根向量**。

据此，矩阵的特征向量便是矩阵的一级根向量，而矩阵的变换矩阵的各个列则是由矩阵的各级根向量组成的。由变换矩阵的非奇异性可知，这些各级根向量应是线性无关的。

由（1.4.3）可知，在矩阵的Jordan标准型中，同一个特征值可能对应着若干个Jordan块。设$\lambda_1,\cdots,\lambda_s$是$A$的相异特征值，$r_i$是$\lambda_i$所对应的Jordan块中的阶数最高的Jordan块的阶，作多项式

$$\psi(\lambda)=(\lambda-\lambda_1)^{r_1}\cdots(\lambda-\lambda_s)^{r_s},\tag{1.4.21}$$

可以证明

$$\psi(A)\equiv(A-\lambda_1 I)^{r_1}\cdots(A-\lambda_s I)^{r_s}=0.\tag{1.4.22}$$

注意，若A的Jordan标准型中，对应某个特征值$\lambda_k(1\leqslant k\leqslant s)$的Jordan块个数大于1，此时对同一特征值的独立的特征向量个数将多于1，$\psi(\lambda)$的次数必将小于矩阵A的阶数n，即必存在A的次数低于n的化零多项式$\psi(\lambda)$。由此，我们给出下述定义。

定义1.4.3 如果矩阵A对应某一特征值λ的线性无关的特征向量多于1时，则称矩阵是**减次**的，否则称为**非减次**的。

亏损与减次、非亏损与非减次是两个独立的概念，但矩阵A既非亏损又非减次的充要条件应是：矩阵A的诸特征值均相异。

§5 特征值扰动问题

5.1 矩阵特征值的连续相依性

设A为n阶方阵，某特征值记为λ，$\tilde{A}$是A的近似矩阵，其特征值记为$\tilde{\lambda}$，且设

$$\tilde{A}=A+\varepsilon B, \tag{1.5.1}$$

式中ε是实的或复的参数，B是元素依模不大于1的矩阵。

所谓矩阵的特征值扰动问题，就是要对$|\lambda-\tilde{\lambda}|$的上界做出尽可能小的估计。在矩阵计算中，由于初始输入的矩阵，往往是某些物理量在仪器的量测精度范围内的测得的结果，或是计算的经舍入处理后的结果，它与真实的物理模型是有差别的。在（1.5.1）中，如把A看作入机数据，$\tilde{A}$看作真实的物理模型，则εB便是它们之间的误差或扰动。可见讨论这种问题是具有普遍意义的。

由于A和$\tilde{A}$的特征值通常各有n个，可见，$|\tilde{\lambda}-\lambda|$不仅与$\varepsilon$有关而且也和$A$、$\tilde{A}$的特征值的排序方法有关，记$A$和$\tilde{A}$的特征多项式为

$$f(\lambda)=\det(\lambda I-A)=\lambda^n+a_1\lambda^{n-1}+\cdots+a_n \tag{1.5.2}$$

$$\varphi(\varepsilon、\lambda)=\det(\lambda I-\tilde{A})=\lambda^n+a_1(\varepsilon)\lambda^{n-1}+\cdots+a_n(\varepsilon) \tag{1.5.3}$$

这里$a_i(i=1,\cdots,n)$是常数，$a_i(\varepsilon)$（$i=1,\cdots,n$）是ε的次数不超过i的多项式，满足

$$\lim_{\varepsilon\to 0}a_i(\varepsilon)=a_i\ (i=1,2,\cdots,n) \tag{1.5.4}$$

定理1.5.1　对任给正数$\sigma>0$，存在$\eta>0$及$f(\lambda)=0$和$\varphi(\varepsilon,\lambda)=0$的根的排序方法，使当$|\varepsilon|<\eta$时，有

$$|\tilde{\lambda}_k-\lambda_k|<(2n-1)\sigma,\ k=1,\cdots,n \tag{1.5.5}$$

其中λ_k、$\tilde{\lambda}_k$分别是$f(\lambda)=0$和$\varphi(\varepsilon,\lambda)=0$的第$k$个根。

证明 设$\lambda_1,\cdots,\lambda_n$是$f(\lambda)=0$的依任意次序排序的$n$个根，当$\lambda_k$是它的$r$重根时，则在$j=1,2,\cdots,n$中有$r$个下标$j_1,\cdots,j_r$使$\lambda_{j_s}=\lambda_k(s=1,\cdots,r)$；对每个$\lambda_k$，以$\lambda_k$为中心以$\sigma$为半径作圆，记为$C_k$。如果当$k\neq j$时$C_k$与$C_j$有重叠部分，则称$C_k$与$C_j$，$\lambda_k$与$\lambda_j$是相邻的；如果对$k\neq j$，可以找到序列：$\lambda_{r_1}$，$\cdots,\lambda_{r_s}$使

$$\lambda_k,\ \lambda_{r_1},\cdots,\lambda_{r_s},\ \lambda_j$$

中的每一个数除λ_k外都与其前一个相邻，称λ_k与λ_j，C_k与C_j是连通的。于是诸圆$C_k(k=1,\cdots,n)$将在λ平面上组成有限个(设为m个）互不连通的连通域，记为：G_1，$G_2,\cdots,G_m$。

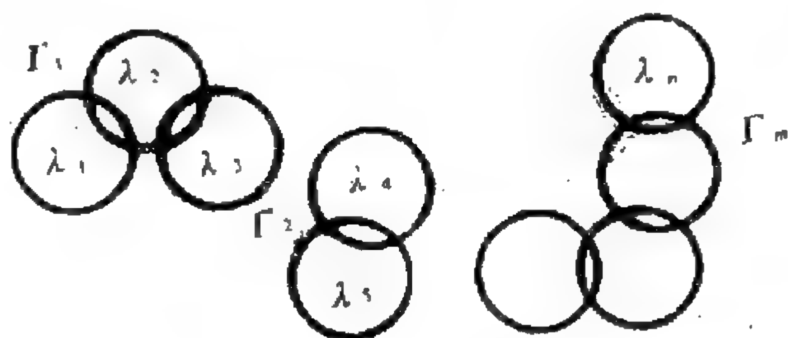


图1.1 σ-连通域

记G_k的边界为Γ_k，它们各由有限个圆弧构成，令

$$\gamma=\max_{z\in\Gamma_k}|z|,\quad 1\leqslant k\leqslant m$$

记
$$f_k=\min_{\lambda\in\Gamma_k}|f(\lambda)|>0,\quad f=\min_{1\leqslant k\leqslant m}(f_k)$$

由于

$$\varphi(\varepsilon,\lambda)=f(\lambda)+\sum_{k=0}^{n-1}(a_k(\varepsilon)-a_k)\lambda^k;$$

当$\lambda\in\Gamma_k$时，由$|\lambda|\leqslant\gamma$，有

$$|\varphi(\varepsilon,\ \lambda)-f(\lambda)|\leqslant\sum_{k=0}^{n-1}|a_k(\varepsilon)-a_k||\lambda|^k.$$

据$a_k(\varepsilon)$的连续性，对任给$\delta>0$，存在$\eta_0>0$，当$|\varepsilon|\leqslant\eta_0$时，

有

$$|\varphi(\varepsilon,\lambda)-f(\lambda)|\leqslant\delta\cdot n\cdot\max(1,\ \gamma^{n-1});$$

当 $\delta<\delta_0=f/(n\cdot\max(1,\ \Gamma^{n-1}))$ 时，知$|\varphi(\varepsilon,\ \lambda)-f(\lambda)|<f\leqslant f_k$。据 Rouche 定理，此时，$\varphi(\varepsilon,\ \lambda)$，$f(\lambda)$ 在每个 Γ_k 围线内应有相同个数的零点；这样，$f(\lambda)$，$\varphi(\varepsilon,\ \lambda)$的零点，即 A、$\tilde{A}$ 的特征值可按下列次序编排：先依 G_k 的次序，从G_1 开始，以任意次序分别对 $f(\lambda)$，$\varphi(\varepsilon,\ \lambda)$ 在围线内部的零点进行编号，之后，再对下一个 G_k 中的各自的零点进行编号，为使符号简化，重编以后的次序，仍记为：$\lambda_1,\ \lambda_2,\cdots,\lambda_n$和 $\tilde{\lambda}_1,\tilde{\lambda}_2,\cdots,\tilde{\lambda}_n$。任取 $1\leqslant k\leqslant n$，设 λ_k，$\tilde{\lambda}_k$ 同属于某个 G_r，且设$f(\lambda)$ 在 G_r 中零点的个数为 $1\leqslant m_r\leqslant n$，因为 λ_k 是G_r 中某个圆的圆心，如果 $\tilde{\lambda}_k$ 所在圆的圆心是通过 s 个圆而与 λ_k 连通起来的，于是

$$|\tilde{\lambda}_k-\tilde{\lambda}_k|\leqslant|\lambda_k-\lambda_{\mu_1}|+|\lambda_{\mu_1}-\lambda_{\mu_2}|+\cdots+|\lambda_{\mu_{s-1}}-\lambda_k|$$
$$\leqslant\sigma+2(s-1)\sigma\leqslant(2n-1)\sigma_\circ$$

由于 σ 的任意性，上述定理可以简单地表述为

$$\lim_{\varepsilon\to 0}\lambda_i(\varepsilon)=\lambda_i\quad(i=1,2,\cdots,n)\qquad(1.5.6)$$

或

$$|\tilde{\lambda}_i(\varepsilon)-\lambda_i|=0\ (|\varepsilon|^{p_i})\ \ (i=1,\cdots,n);\qquad(1.5.7)$$

其中$p_i>0$，而 $\lambda_1,\cdots,\lambda_n$，$\lambda_1(\varepsilon),\cdots,\lambda_n(\varepsilon)$ 各对应于A、$A+\varepsilon B$ 的特征值和某一排序方法。

定义1.5.1 在（1.5.7）中，当 $p_i\geqslant 1$ 时，称矩阵A的特征值问题在 λ_i 处是**良态**的，否则称为**病态**的。

例1.5.1 设

$$A=\begin{pmatrix}\lambda & 1 & & \\ & \lambda & 1 & \\ & & \ddots & \ddots \\ & & & 1 \\ & & & \lambda\end{pmatrix},$$

矩阵B仅在（n，1）元素处为1，其余均为零，此时，由

$$\det(\tilde{\lambda}I - A - \varepsilon B) = (\tilde{\lambda} - \lambda)^n + (-1)^n\varepsilon = 0,$$

得
$$|\tilde{\lambda} - \lambda| = |\varepsilon|^{\frac{1}{n}},$$

即 $p = \frac{1}{n}(<1)$，A 的特征值问题是病态的.

5.2 Gerschgorin 定理

为了对矩阵特征值的性态问题做进一步的研究，本段将以 Gerschgorin 圆盘定理及对角相似变换技术为工具来讨论这一问题.

设 n 阶方阵

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & & a_{nn} \end{pmatrix}, \tag{1.5.8}$$

对每个 $1 \leqslant i \leqslant n$，称

$$C_i = \{\lambda \mid |\lambda - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}| = \rho_i\} \tag{1.5.9}$$

为 A 的第 i 个 (Gerschgorin) 圆盘，这里 a_{ij} 可为实数或复数.

定理1.5.2 A 的任意一个特征值 λ 至少落在 A 的一个圆盘上.

证明 设 x 是 A 的对应于 λ 的特征向量，且设其依模最大的分量为 $x_i = 1$，据 $Ax = \lambda x$ 的分量形式，有

$$\sum_{j=1}^{n} a_{kj}x_j = \lambda x_k, \quad k = 1, 2, \cdots, n$$

当 $k = i$ 时，得

$$\lambda - a_{ii} = \sum_{j \neq i} a_{ij}x_j,$$

故

$$|\lambda - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}|. \tag{1.5.10}$$

定理1.5.3 设 A 的 n 个圆盘中，有 s 个圆盘组成一个连

通区域 S，且与其余的 $n-s$ 个圆盘互不相交，则 A 的 n 个特征值中，有且仅有 s 个落在 S 上。

证明 令 $A=D+C$，其中 D 是 A 的对角阵。作

$$A(\varepsilon)=D+\varepsilon C,$$

当 ε 由 0 增加到 1 时，$A(\varepsilon)$ 由 D 变到 A，显然 $A(0)$ 的特征值就是 D 的对角元，它们分别是 n 个圆盘的圆心，其中有 s 个圆心在 S 上，由于 $A(\varepsilon)$ 的特征值连续依赖于 ε，所以，当 ε 由 0 增加到 1 时，$A(\varepsilon)$ 的 n 个特征值将在 λ 平面上由各自的圆心出发画出 n 条连续曲线，这些曲线中的每一条或许全部落在 S 上，或许全部落在其余的 $n-s$ 个圆盘的和集上，由 S 上的 s 个圆的圆心出发的特征值，在 $0<\varepsilon<1$ 时应全部落在 S 上，而此种特征值有且仅有 s 个。

推论1.5.1 设 A 有一个圆盘 C_i，它孤立于其余的圆盘，则 $A(\varepsilon)$ 在此圆盘内有且仅有一个特征值 $\tilde{\lambda}(\varepsilon)$，当以 a_{ii} 作为 $\tilde{\lambda}(\varepsilon)$ 的近似值时，误差将不超过 $|\varepsilon|\sum\limits_{j\neq i}|a_{ij}|$，即

$$|\tilde{\lambda}(\varepsilon)-a_{ii}|\leqslant\varepsilon\sum_{j\neq i}|a_{ij}|. \tag{1.5.11}$$

下面我们将利用上述定理和推论及对角相似变换技术来讨论矩阵特征值问题的性态。

不失一般性，设 λ_1 是任意矩阵 A 的一个特征值。先就 λ_1 是 A 的单重特征值的情形进行讨论。为方便起见，进一步再假设 A 是非亏损的矩阵。

我们的问题是要研究 A 的带扰动的矩阵 $\tilde{A}=A+\varepsilon B$ 的某一特征值 $\tilde{\lambda}_1(\varepsilon)$ 与 λ_1 的关系，这里假定当 $\varepsilon\to 0$ 时 $\tilde{\lambda}_1(\varepsilon)\to\lambda_1$。此时，由于存在非奇异矩阵 H 使

$$AH=H\mathrm{diag}(\lambda_1,\cdots,\lambda_n), \tag{1.5.12}$$

其中 $\lambda_1\neq\lambda_i(i\geqslant 2)$，于是

$$H^{-1}(A+\varepsilon B)H=\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)+\varepsilon H^{-1}BH$$

$$=\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n\end{pmatrix}+\varepsilon\begin{pmatrix}\beta_{11}/s_1 & \beta_{12}/s_1 & \cdots & \beta_{1n}/s_1 \\ \beta_{21}/s_2 & \beta_{22}/s_2 & \cdots & \beta_{2n}/s_2 \\ \vdots & \vdots & & \vdots \\ \beta_{n1}/s_n & \beta_{n2}/s_n & \cdots & \beta_{nn}/s_n\end{pmatrix}. \tag{1.5.13}$$

这里，我们假定H是由A的完全标准化特征向量系组成的，即

$$H=(x_1, \cdots, x_n),\quad \|x_i\|=1\ (1\leqslant i\leqslant n),$$
$$H^{-T}=(y_1/s_1, \cdots, y_n/s_n),\quad \|y_i\|=1\ (1\leqslant i\leqslant n), \tag{1.5.14}$$
$$s_i=y_i^T x_i\neq 0,\quad i=1, \cdots, n,$$
$$y_i^T x_j=0,\quad i\neq j;$$

而

$$\beta_{ij}=y_i^T B x_j,\quad i,\ j=1, \cdots, n. \tag{1.5.15}$$

由于$\lambda_1\neq\lambda_i$ $(i=2, \cdots, n)$，故（1.5.13）右端的第1个圆盘，当ε充分小时，必将孤立于其余的圆盘，此时若以

$$\tilde{\lambda}_1(\varepsilon)=\lambda_1+\varepsilon\beta_{11}/s_1 \tag{1.5.16}$$

作为$A+\varepsilon B$的一个近似特征值，则误差将不超过

$$\rho(\varepsilon)=|\varepsilon|\sum_{j\neq 1}|\beta_{1j}/s_1|=O(|\varepsilon|),$$

但这一估计还可进一步精确化。为此，用D^{-1}和D分别左乘和右乘（1.5.13）的两端，其中取

$$D=\mathrm{diag}(k/\varepsilon, 1, \cdots, 1),$$

于是

$$D^{-1}H^{-1}(A+\varepsilon B)HD$$

$$=\mathrm{diag}(\lambda_i)+\varepsilon\begin{pmatrix}\beta_{11}/s_1 & \varepsilon\beta_{12}/ks_1 & \cdots & \varepsilon\beta_{1n}/ks_1 \\ k\beta_{21}/\varepsilon s_2 & \beta_{22}/s_2 & \cdots & \beta_{2n}/s_2 \\ \vdots & \vdots & & \vdots \\ k\beta_{n1}/\varepsilon s_n & \beta_{n2}/s_n & \cdots & \beta_{nn}/s_n\end{pmatrix}. \tag{1.5.17}$$

此时，第1个和第i个圆盘的中心和半径为：

圆心　　　　半径

第 1 个　$\lambda_1+\varepsilon\beta_{11}/s_1$　　$(\varepsilon^2/|ks_1|)\sum_{j\neq 1}|\beta_{1j}|$；

第 i 个　$\lambda_i+\varepsilon\beta_{ii}/s_i$　　$|k\beta_{i1}/s_i|+\left|\frac{\varepsilon}{s_i}\right|\sum_{j\neq 1,i}|\beta_{ij}|$。

欲使第 1 个圆盘孤立于其余的圆盘，只需

$$\left|\lambda_1-\lambda_i+\varepsilon\left(\frac{\beta_{11}}{s_1}-\frac{\beta_{ii}}{s_i}\right)\right|>|k\beta_{i1}/s_i|+\left|\frac{\varepsilon}{s_i}\right|\sum_{j\neq 1,i}|\beta_{ij}|$$

$$+\left|\frac{\varepsilon^2}{ks_1}\right|\sum_{j\neq 1}|\beta_{1j}|.$$

为此，可先取 k 使　$\max\limits_{2\leqslant i\leqslant n}|k\beta_{i1}/s_i|<\frac{1}{2}\min\limits_{i\geqslant 2}|\lambda_1-\lambda_i|$，

再取 ε 使

$$|\varepsilon|\left(\left|\frac{\beta_{11}}{s_1}\right|+\left|\frac{\beta_{ii}}{s_i}\right|+\left|\frac{1}{s_i}\right|\sum_{j\neq 1,i}|\beta_{ij}|\right)+\left|\frac{\varepsilon}{ks_1}\right|\sum_{j\neq 1}|\beta_{1j}|$$

$$<\frac{1}{2}\min_{i\geqslant 2}|\lambda_1-\lambda_i|.$$

据（1.5.17），及 Gerschgorin 定理，则有

$$|\tilde{\lambda}_1(\varepsilon)-\tilde{\lambda}_1'(\varepsilon)|\leqslant\left|\frac{\varepsilon^2}{ks_1}\right|\sum_{j\neq 1}|\beta_{1j}|=O(|\varepsilon|^2). \qquad (1.5.18)$$

当 λ_1 是 A 的单重特征值但 A 本身是亏损矩阵时，在 A 的 Jordan 标准型中，对应于其余的特征值 $\lambda_i(i\geqslant 2)$ 将有阶数大于 1 的 Jordan 子块，此时，仿照上述的讨论过程，仍可得到（1.5.18）中的结果。

当 A 的 Jordan 标准型中，对应于 λ_1 的Jordan 块中含有阶数大于 1 的块时，设相应的最高阶数为 r_1，此时，只能得到

$$|\tilde{\lambda}_1(\varepsilon)-\lambda_1|=O(|\varepsilon|^{\frac{1}{r_1}}) \qquad (1.5.19)$$

的估计，即相应的特征值扰动问题是病态的。

§6 特征值问题的条件数

6.1 谱条件数

本节所讨论的矩阵A均指非亏损矩阵。

设λ_1，$\lambda_2, \cdots, \lambda_n$为$A$的特征值，$\boldsymbol{x}_1$，$\boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$是相应的完全标准特征向量系。

$$A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i \quad (i = 1, \cdots, n), \tag{1.6.1}$$

记

$$H = (\boldsymbol{x}_1, \ \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n), \tag{1.6.2}$$

令

$$H^{-T} = (\boldsymbol{y}_1/s_1, \cdots, \boldsymbol{y}_n/s_n), \tag{1.6.3}$$

其中$\|\boldsymbol{y}_i\| = 1$，且

$$\boldsymbol{y}_j^T \boldsymbol{x}_i = \begin{cases} s_i \neq 0 & (i = j); \\ 0 & (i \neq j). \end{cases} \tag{1.6.4}$$

设$\tilde{\lambda}(\varepsilon)$是$A + \varepsilon B$的一个特征值，则$A + \varepsilon B - \tilde{\lambda} I$是奇异阵，知

$$H^{-1}(A + \varepsilon B - \tilde{\lambda} I) H = \mathrm{diag}(\lambda_i - \tilde{\lambda}) + \varepsilon H^{-1} B H$$

亦应是奇异阵，当$\lambda_i \neq \tilde{\lambda} (i = 1, \cdots, n)$时，由于

$$\begin{aligned} & H^{-1}(A + \varepsilon B - \tilde{\lambda} I) H \\ & = \mathrm{diag}(\lambda_i - \tilde{\lambda}) [I - \varepsilon (\mathrm{diag}(\lambda_i - \tilde{\lambda}))^{-1} H^{-1} B H] \end{aligned}$$

可知

$$\|\varepsilon (\mathrm{diag}(\lambda_i - \tilde{\lambda}))^{-1} H^{-1} B H\| \geqslant 1 \tag{1.6.5}$$

对任意意义的矩阵范数成立，而在欧氏范数意义下，有

$$|\varepsilon| \max_i (|\lambda_i - \tilde{\lambda}|^{-1}) \|H^{-1}\| \|H\| \|B\| \geqslant 1$$

或

$$\min_i |\lambda_i - \tilde{\lambda}| \leqslant \|\varepsilon\| \|H^{-1}\| \|H\| \|B\|, \tag{1.6.6}$$

上式对 $\lambda_i=\tilde{\lambda}$ 也是成立的，故在原讨论中关于 $\lambda_i\neq\tilde{\lambda}$ 的限制可以去掉。记

$$\kappa(H)=\|H^{-1}\|\|H\|, \tag{1.6.7}$$

可知

$$|\lambda_i-\tilde{\lambda}|\leqslant|\varepsilon|\kappa(H)\|B\| \tag{1.6.8}$$

至少对 i 的一个值成立。

注意，使得 $H^{-1}AH=\mathrm{diag}(\lambda_1,\cdots,\lambda_n)$ 的 H 一般是非唯一的，如令

$$\kappa(A)=\inf_H(\kappa(H)), \tag{1.6.9}$$

显然，$\kappa(A)$ 能为某个 H 所取到，且类似地有

$$|\lambda_i-\tilde{\lambda}|\leqslant|\varepsilon|\kappa(A)\|B\|. \tag{1.6.10}$$

由于 $\kappa(A)$ 是用矩阵的谱范数定义的，且因 $k(A)$ 的大小能刻划 A 的特征值扰动的敏感性，因此，通常称 $\kappa(A)$ 为 A 的对特征值问题来说的谱条件数。

当 $\kappa(A)$ 较小时，人们有时也把 A 的特征值问题叫做是良态的，而把相反的情形叫做是病态的。但此地所谓的良态、病态与本章第 5 节定义的良态与病态的概念实际上是有区别的。由于本节只讨论非亏损阵的情形，而非亏损阵的特征值问题在第 5 节中均应算作是良态的。因此，本节中提到的良态、病态概念实际是上一节中的良态概念的更细致的区分。关于谱条件数 $\kappa(A)$，可以证明它具有以下一些简单的性质：

(1) $\kappa(A)\geqslant 1$, (1.6.11)

(2) 当 A 是正规阵时，$\kappa(A)=1$. (1.6.12)

由 (1.6.8) 还可建立类似于定理 1.5.3 的定理。

定理1.6.1 如果圆盘

$$|\lambda-\lambda_i|\leqslant|\varepsilon|\kappa(A)\|B\|,\quad i=1,2,\cdots,n \tag{1.6.13}$$

中有 s 个圆盘的和集组成与其余 $n-s$ 个圆盘互不连通的连通域

S，则 $A+\varepsilon B$ 的特征值中恰有 s 个落在 S 上。

6.2 n-条件数

当λ_1是非亏损矩阵A的单重特征值时，据（1.5.16）及（1.5.17）知道，当以

$$\lambda_1'(\varepsilon)=\lambda_1+\varepsilon\beta_{11}/s_1 \tag{1.6.14}$$

作为 $A+\varepsilon B$ 的一个特征值 $\lambda_1(\varepsilon)$的近似值时，有

$$\lambda_1(\varepsilon)-\lambda_1=\varepsilon\beta_{11}/s_1+O(\varepsilon^2). \tag{1.6.15}$$

当$|s_1|$很小时，上式表明相应的特征扰动对于矩阵元素的扰动将是十分敏感的。据此，我们一般称

$$1/s_i, i=1,2,\cdots,n \tag{1.6.16}$$

为矩阵A的 n-条件数。诸s_i 的计算可依（1.4.10）~（1.4.12）的公式进行。

谱条件数只能从总体上讨论特征扰动的敏感性而不能按特征值逐个地进行讨论，因此，n-条件数的引入意在补救其不足之处。

6.3 n-条件数与谱条件数的关系

设H是（1.6.9）右端中取得准确下界的阵，

$$x_i=He_i/\|He_i\|,\quad y_i=H^{-T}e_i/\|H^{-T}e_i\|,$$

于是

$$|s_i|=|y_i^Tx_i|=\frac{|e_i^TH^{-1}He_i|}{\|He_i\|\|H^{-T}e_i\|}$$

$$=\frac{1}{\|He_i\|\|H^{-T}e_i\|}\geqslant\frac{1}{\kappa(A)},$$

即

$$\left|\frac{1}{s_i}\right|\leqslant\kappa(A),\quad i=1,2,\cdots,n. \tag{1.6.17}$$

另一方面，由

$$H^{-1}AH=\mathrm{diag}(\lambda_1,\cdots,\lambda_n),$$

$$H=(x_1,\cdots,x_n),\quad H^{-T}=\left(\frac{y_1}{s_1},\cdots,\frac{y_n}{s_n}\right),$$

记

$$D=\text{diag}(|s_1|^{-\frac{1}{2}},\cdots,|s_n|^{-\frac{1}{2}}),$$

有

$$(HD)^{-1}A(HD)=\text{diag}(\lambda_1,\cdots,\lambda_n),$$

据

$$\kappa(A)\leqslant\|HD\|\|(HD)^{-1}\|\leqslant\|HD\|_E\|(HD)^{-1}\|_E$$

$$\leqslant\left(\sum_{i=1}^{n}\frac{\|x_i\|^2}{|s_i|}\right)^{\frac{1}{2}}\left(\sum_{i=0}^{n}\frac{\|y_i^T\|^2}{|s_i|}\right)^{\frac{1}{2}}=\sum_{i=1}^{n}\frac{1}{|s_i|},\quad(1,6,18)$$

综合（1.6.15）和（1.6.16），有

$$\frac{1}{|s_i|}\leqslant\kappa(A)\leqslant\sum_{i=1}^{n}\frac{1}{|s_i|}.\qquad(1.6.19)$$

这个不等式说明：当A的某一特征值属于病态时，则就总体上来说的特征值问题必是病态的，而当总体的特征值问题属于病态时，则在个别情况下的特征问题中至少有一个是病态的。

谱条件数和n-条件数还有一个重要性质，即二者在酉相似变换下是不变的。

设R是酉阵，且$B=RAR^H$，若$H^{-1}AH=\text{diag}(\lambda_1,\cdots,\lambda_n)$，则有 $(RH)^{-1}B(RH)=\text{diag}(\lambda_1,\cdots,\lambda_n)$，记$B$的谱条件数$\kappa'$，则

$$\kappa'=\inf_H(\|(RH)^{-1}\|_2\|RH\|_2)$$

$$=\inf_H(\|H^{-1}\|_2\|H\|_2)=\kappa.\qquad(1.6.20)$$

若以x_i，y_i表示A的右、左特征向量，则B的右、左特征向量为：

$$x_i'=Rx_i,\ \ y_i'^T=e_i^T(RH)^{-1}/\|e_i(RH)^{-1}\|_2=y_i^TR^H$$

于是

$$s_i'=y_i'^Tx_i'=y_i^TR^HRx_i=y_i^Tx_i=s_i.\qquad(1.6.21)$$

$$(i=1,2,\cdots,n)$$

§7 实对称矩阵的扰动

7.1 实对称扰动

实对称阵、Hermite 阵及正规阵之间存在着前者包含于后者的关系，且均属于非亏损的矩阵，它们的特征扰动问题必有更深入的结果。本节讨论实对称阵的性质及其特征扰动问题，有关的结论常可推之于后面的两类矩阵，但在叙述时将不一一指出。

设A是实对称阵，其特征值为$\lambda_i(i=1,2,\cdots,n)$，它们对应的右的和左的标准化特征向量设为$x_i$和$y_i$；不论诸$\lambda_i$是否相异，可取$x_i=y_i$（$i=1,\cdots,n$）为实分量的向量，于是有

$$s_i=y_i^T x_i=1 \quad (i=1,\cdots,n), \tag{1.7.1}$$

因此，实对称阵的特征值扰动问题总是良态的。记$\lambda(\varepsilon)$为$A+\varepsilon B$的任意一个特征值，仿本章§6(1.6.1)～(1.6.8)的证明，可知

$$|\lambda(\varepsilon)-\lambda_i|\leqslant|\varepsilon|\|B\|\leqslant|\varepsilon|\|B\|_E\leqslant n|\varepsilon| \tag{1.7.2}$$

至少对某一$i=1,2,\cdots,n$成立。即$A+\varepsilon B$的任意一个特征值必将落在某一个以λ_i为中心以$n|\varepsilon|$为半径的圆盘上。由此，可以导出一个有趣的结论：如果λ_i是A的单重实特征值，则当ε充分小时，只要εB是实的但不一定对称，$A+\varepsilon B$必有一个靠近λ_i的实特征值。

因为λ_i既是实数又是单重特征值，于是可以假设A有一个异于λ_i又最靠近λ_i的特征值λ_j，当$|\varepsilon|<|\lambda_i-\lambda_j|/2n$时，设$\lambda(\varepsilon)$是$A+\varepsilon B$的且落在$|\lambda-\lambda_i|<n|\varepsilon|$圆盘上的特征值，当$k\neq i$时，由

$$|\lambda_k-\lambda(\varepsilon)|\geqslant|\lambda_k-\lambda_i|-|\lambda_i-\lambda(\varepsilon)|>n|\varepsilon|. \tag{1.7.3}$$

即第 i 个圆盘将孤立于其余的圆盘，此时 $\lambda(\varepsilon)$ 必为实数，不然，则因 $\bar{\lambda}(\varepsilon)$ 也将是 $A+\varepsilon B$ 的特征值，且也落在第 i 个圆盘上，这将与第 i 个圆盘上只能有 $A+\varepsilon B$ 的一个特征值的推理矛盾。

此外，如果 A 的特征值均相异，则当 $|\varepsilon|$ 充分小且 εB 为实阵时，$A+\varepsilon B$ 的特征值亦将全是实数且相异，此时 $A+\varepsilon B$ 应有完全的特征向量系。当 B 不仅是实的且也是对称阵时，不论实参数 ε 的大小如何，扰动后的矩阵仍是实对称矩阵，在下面的讨论中，我们将把 ε 取为 1，并把 $|B|\leqslant 1$ 的限制去掉，而提出更一般的问题。

7.2 实对称阵的和的特征值

设 A、B 均为实对称阵，记

$$C=A+B. \tag{1.7.4}$$

下面分两种情况进行讨论。

7.2.1 C 是对角加边的实对称阵

$$C=\begin{pmatrix} a & \vdots & \boldsymbol{a}^T \\ \cdots & \cdots & \cdots \\ \boldsymbol{a} & \vdots & \mathrm{diag}(a_i) \end{pmatrix}. \tag{1.7.5}$$

定义1.7.1 设 $x_1, x_2, \cdots, x_n, x_{n+1}$ 及 $y_1, \cdots, y_n$ 为两组实数，当

$$x_1>y_1>x_2>y_2>\cdots x_n>y_n>x_{n+1} \tag{1.7.6}$$

时，称这两组数在严格意义下是**隔离**的。当 (1.7.6) 的“$>$”被换为“$\geqslant$”时，称它们在弱的意义下是隔离的。

定理1.7.1 当 (1.7.5) 中的 a_i 均相异且 $\boldsymbol{a}$ 的分量均不为零时，则 C 的特征值：$\lambda_1>\lambda_2>\cdots>\lambda_n$ 与诸 a_i 在严格的意义下是隔离的。

证明 因 C 经行列交换的相似变换后仍具有 (1.7.5) 的形状，且特征值不变，故可设 (1.7.5) 中的 a_i 满足：$a_1>a_2$

$>\cdots>\alpha_{n-1}$，此时C的特征方程为

$$(\alpha-\lambda)\prod_{i=1}^{n-1}(\alpha_i-\lambda)-\sum_{i=1}^{n-1}a_i^2\prod_{\substack{j\neq i\\ j=1}}^{n-1}(\alpha_j-\lambda)=0, \tag{1.7.7}$$

记

$$f(\lambda)=\lambda+\sum_{i=1}^{n-1}a_i^2/(\alpha_i-\lambda), \tag{1.7.8}$$

则（1.7.7）化为

$$\alpha-\lambda-\sum_{i=1}^{n-1}a_i^2/(\alpha_i-\lambda)=\alpha-f(\lambda)=0. \tag{1.7.9}$$

如图1.2，因

$$f'(\lambda)=1+\sum_{i=1}^{n-1}a_i^2/(\alpha_i-\lambda)^2>0, \tag{1.7.10}$$

$$\alpha_{i+1}<\lambda<\alpha_i,\quad i=1,2,\cdots,n-2.$$

方程（1.7.9）在（α_{i+1}，α_i）内有根，记为λ_{i+1}，在（$-\infty$，α_n）和（α_1，∞）的根记为λ_n和λ_1则λ_1，$\lambda_2,\cdots,\lambda_n$与$\alpha_1,\cdots,\alpha_{n-1}$，在严格的意义下是隔离的，即$\lambda_1>\alpha_1>\lambda_2>\cdots>\alpha_{n-1}>\lambda_n$。

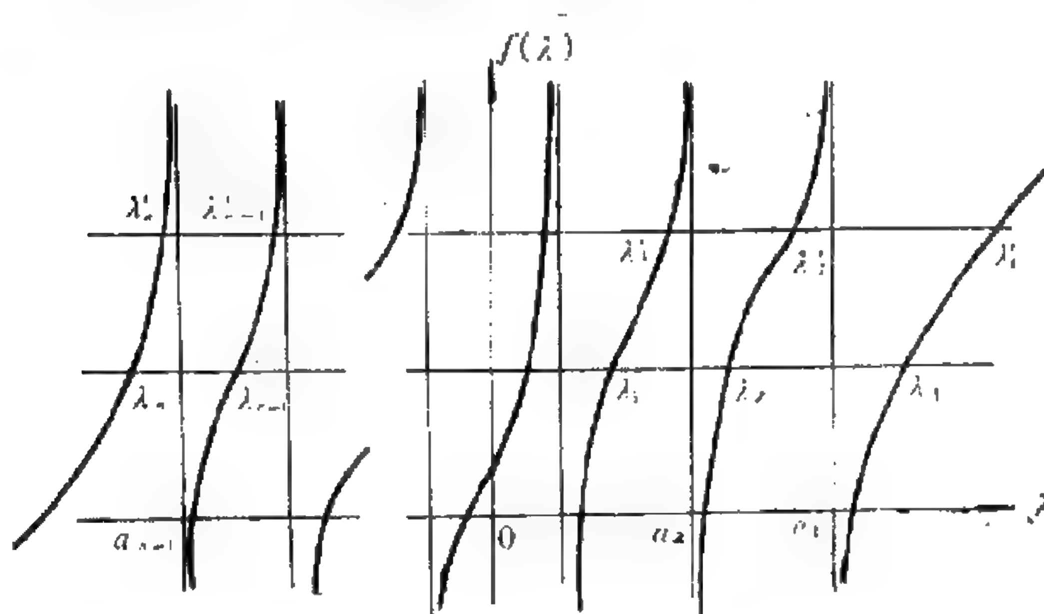


图1.2　特征值的隔离

当某个α_i是$\mathrm{diag}(\alpha_1,\cdots,\alpha_n)$的$r_i$重特征值时，由

$$\alpha_{i-1}>\alpha_i=\alpha_{i+1}=\cdots=\alpha_{i+r_{i-1}}>\alpha_{i+r_i}$$

可见

$$\alpha_i=\lambda_{i+1}=\alpha_{i+1}=\cdots=\lambda_{i-r_i-1}=\alpha_{r_i+i-1},$$

即 λ_{i+1} 必为 C 的 r_{i-1} 重特征值，此时，$\lambda_1,\cdots,\lambda_n$ 与 $\alpha_1,\cdots,\alpha_{n-1}$ 在弱的意义下是隔离的。

当 $\boldsymbol{a}$ 的某些分量是零时，则与之同行的 α_i 必为 C 的特征值，这些 α_i 将既属于 C 的特征值数组，又属数组 $\alpha_1,\cdots,\alpha_{n-1}$，把 C 中这些 α_i 所在的行列去掉，据前面的证明知，剩下的矩阵的特征值将在弱的意义下为 $\mathrm{diag}(\alpha_i)$ 的剩下的那些 α_i 所隔离。而在弱的意义下互相隔离的两个数组中，各自分别插入一些相同的数后，仍将在弱的意义下相互隔离。于是，不管诸 α_i 是否相异，也不管 $\boldsymbol{a}$ 是否有零分量，C 的特征值将在弱的意义下为 $\alpha_1,\cdots,\alpha_{n-1}$ 所隔离。

当 $\boldsymbol{a}$ 及诸 α_i 均不变，而 α 从 α 变到 α' 时，设 C 的特征值由 λ_i 变到 λ_i'，而 λ_i 与 α_i 的隔离情况与 λ_i' 和 α_i 的隔离情况是一样的。因此，λ_i 中的重特征值与 λ_i' 中的重特征值将不变化。当 λ_i、λ_i' 都是单重特征值时，由于

$$f(\lambda'_k)-f(\lambda_k)=\lambda_k'+\sum_{i=1}^{n-1}a_i^2/(\alpha_i-\lambda_i')-\lambda_k-\sum_{i=1}^{n-1}a_i^2/(\alpha_i-\lambda_i)$$

$$=\alpha'-\alpha, \tag{1.7.11}$$

有

$$(\lambda_k'-\lambda_k)f'(\xi_k)=\alpha'-\alpha,\quad \lambda_k'-\lambda_k=m_k(\alpha'-\alpha). \tag{1.7.12}$$

这里 ξ_k 在 λ_k 与 λ_k' 之间，$0<m_k=1/f'(\xi_k)<1$。由 (1.7.7) 及根与系数的关系，得

$$\sum\lambda_k=\alpha+\sum\alpha_i,\quad \sum\lambda_k'=\alpha'+\sum\alpha_i,$$

故

$$\sum\lambda_k'-\sum\lambda_k=\alpha'-\alpha,$$

从而

$$\sum m_k=1. \tag{1.7.13}$$

由（1.7.12）和（1.7.13）可知对角加边矩阵 C 中，当 α 发生 $\Delta\alpha=\alpha'-\alpha$ 的改变量时，特征值移动的方向与 $\Delta\alpha$ 的方向相同，且特征值的总变化量等于 $\Delta\alpha$。

7.2.2　$C=A+B$，B 是秩1矩阵。

因为 B 是秩 1 的，存在正交阵 R 使

$$B=R\begin{pmatrix}\rho & \vdots & 0\\ \cdots & \cdots & \cdots\\ 0 & \vdots & 0\end{pmatrix}R^T。\qquad(1.7.14)$$

在 $C=A+B$ 的两端分别用 R、R^T 左乘和右乘，记

$$RAR^T=\begin{pmatrix}\alpha & \vdots & \boldsymbol{a}^T\\ \cdots & \cdots & \cdots\\ \boldsymbol{a} & \vdots & A_{n-1}\end{pmatrix}.$$

这里 A_{n-1} 是 $n-1$ 阶的实对称阵，得

$$RCR^T=\begin{pmatrix}\alpha+\rho & \vdots & \boldsymbol{a}^T\\ \cdots & \cdots & \cdots\\ \boldsymbol{a} & \vdots & A_{n-1}\end{pmatrix}.$$

对 A_{n-1}，存在 $n-1$ 阶实正交阵 S，使

$$SA_{n-1}S^T=\mathrm{diag}(\alpha_1,\cdots,\alpha_{n-1}),$$

作

$$Q=\begin{pmatrix}1 & \vdots & 0\\ \cdots & \cdots & \cdots\\ 0 & \vdots & S\end{pmatrix}R$$

则

$$Q(A+B)Q^T=\begin{pmatrix}\alpha+\rho & \vdots & \boldsymbol{b}^T\\ \cdots & \cdots & \cdots\\ \boldsymbol{b} & \vdots & \mathrm{diag}(\alpha_i)\end{pmatrix},$$

这里 $\boldsymbol{b}=S\boldsymbol{a}$，上式右端已化为对角加边的实对称阵，由于 A 和 $A+B$ 的特征值就是

$$\begin{pmatrix}\alpha & \vdots & \boldsymbol{b}^T\\ \cdots & \cdots & \cdots\\ \boldsymbol{b} & \vdots & \mathrm{diag}(\alpha_i)\end{pmatrix}\text{和}\begin{pmatrix}\alpha+\rho & \vdots & \boldsymbol{b}^T\\ \cdots & \cdots & \cdots\\ \boldsymbol{b} & \vdots & \mathrm{diag}(\alpha_i)\end{pmatrix}$$

的特征值，设其各为λ_i和λ_i'，据（1.7.12），有

$$\begin{aligned}&\lambda_i' - \lambda_i = m_i\rho,\\&\sum m_i = 1 \quad (0 \leqslant m_i \leqslant 1).\end{aligned} \tag{1.7.15}$$

注意，由于ρ是B的唯一的非零特征值，上式表明：当对称阵的扰动阵是秩1的对称阵时，则特征值的扰动与扰动阵的唯一的非零特征值同号，且特征值扰动之和等于扰动阵的特征值。

§8 Hermite 阵的极值性质

8.1 Hermite 阵的极值性质

本节设A为 Hermite 阵，其特征值是实数：

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n. \tag{1.8.1}$$

存在酉阵

$$U = (x_1, x_2, \cdots, x_n) \tag{1.8.2}$$

使

$$U^H A U = \mathrm{diag}(\lambda_1, \cdots, \lambda_n). \tag{1.8.3}$$

视x_i $(i = 1, \cdots, n)$ 为n维复向量空间C^n的一组标准正交基。对任意的$x \in C^n$，x可以分别表为

$$\begin{aligned}&x = \xi_1 e_1 + \xi_2 e_2 + \cdots + \xi_n e_n;\\&x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n.\end{aligned} \tag{1.8.4}$$

当$x^H x = 1$时，可得

$$\begin{aligned}x^H x &= |\xi_1|^2 + |\xi_2|^2 + \cdots + |\xi_n|^2\\&= |\alpha_1|^2 + |\alpha_2|^2 + \cdots + |\alpha_n|^2 = 1.\end{aligned} \tag{1.8.5}$$

由于

$$f(x) = x^H A x = \sum_{i,j=1}^{n} a_{ij}\bar{\xi}_i \xi_j, \tag{1.8.6}$$

$$a_{ij} = \bar{a}_{ij} \quad (i, j = 1, \cdots, n),$$

易知 $f(\boldsymbol{x})$ 是 $\mathbf{C}^n$ 上的实的连续泛函，它在球面

$$\boldsymbol{x}^H\boldsymbol{x}=1 \tag{1.8.7}$$

取得最大值

$$\rho_1=\max_{\|x\|=1}(\boldsymbol{x}^H A\boldsymbol{x})$$

由于

$$\boldsymbol{x}^H A\boldsymbol{x}=\sum_{i=1}^{n}\lambda_i|\alpha_i|^2\leqslant\lambda_1\|\boldsymbol{x}\|^2=\lambda_1,$$

故

$$\rho_1\leqslant\lambda_1.$$

另外，由

$$\boldsymbol{x}_1^H A\boldsymbol{x}_1=\lambda_1\|\boldsymbol{x}\|^2=\lambda_1,$$

又有

$$\rho_1\geqslant\lambda_1,$$

得

$$\rho_1=\max_{\|x\|=1}(\boldsymbol{x}^H A\boldsymbol{x})=\lambda_1. \tag{1.8.8}$$

当对 x 不仅给以 $\boldsymbol{x}^H\boldsymbol{x}=1$ 的约束，而且给以 $\boldsymbol{x}_1^H\boldsymbol{x}=0$的约束时，可证

$$\max_{\substack{\|x\|=1\\ x_1^H x=0}}(\boldsymbol{x}^H A\boldsymbol{x})=\lambda_2. \tag{1.8.9}$$

一般地，可证：

$$\max_{\substack{\|x\|=1\\ x_i^H x=0\\ i=1,\cdots,k}}(\boldsymbol{x}^H A\boldsymbol{x})=\lambda_k \qquad (1\leqslant k+1\leqslant n). \tag{1.8.10}$$

在上述讨论中，当要确定 A 的第 $i+1$ 个特征值时，应先知道对应于 A 的前 i 个特征值 $\lambda_1,\cdots\lambda_i$ 的那些特征向量 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i$，而利用下面的讨论则可避免这一点。

今考虑求 $\boldsymbol{x}^H A\boldsymbol{x}$ 的极小极大极小问题：

$$\min_{P_i} \max_{\substack{\|x\|=1 \\ P_i^H x=0, P_i \neq 0 \\ i=1,\cdots,k}} (x^H A x), \tag{1.8.11}$$

这里 $P_1, \cdots, P_k$ 为任意的非零向量。仍令 $U=(x_1, \cdots, x_n)$，且记 $x=Uy$，则（1.8.11）化为

$$\min_{q_i} \max_{\substack{\|y\|=1 \\ q_i^H y=0, q_i \neq 0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y. \tag{1.8.12}$$

这里

$$q_i = U^H P_i, \quad i=1,2,\cdots,k. \tag{1.8.13}$$

因为对于问题（1.8.12），存在一组向量 $\bar{q}_i (i=1,\cdots,k)$ 使

$$\min_{q_i} \max_{\substack{\|y\|=1 \\ q_i^H y=0, q_i \neq 0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y = \max_{\substack{\|y\|=1 \\ \bar{q}_i^H y=0, \bar{q}_i \neq 0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y$$

$$\geqslant \max_{\substack{\|y\|=1, \quad \eta_1=\cdots=\eta_k=0 \\ \bar{q}_i^H y=0, \quad \bar{q}_i \neq 0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y = \max_{\substack{\|y\|=1 \\ \bar{q}_i^H y=0, \bar{q}_i \neq 0 \\ i=1,\cdots,k}} \sum_{i=k+1}^{n} \lambda_i |\eta_i|^2 \geqslant \lambda_{k+1}.$$

这里 η_i 是 y 的第 i 个分量。另一方面，

$$\min_{q_i} \max_{\substack{\|y\|=1 \\ q_i^H y=0, q_i \neq 0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y \leqslant \max_{\substack{\|y\|=1 \\ e_i^T y=0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y$$

$$= \sum_{i=k+1}^{n} \lambda_i |\eta_i|^2 \leqslant \lambda_{k+1},$$

最后得

$$\min_{q_i} \max_{\substack{\|y\|=1 \\ q_i^H y=0, q_i \neq 0 \\ i=1,\cdots,k}} y^H \mathrm{diag}(\lambda_i) y = = \lambda_{k+1},$$

即

$$\min_{P_i} \max_{\substack{\|x\|=1 \\ P_i^H x=0, P_i \neq 0 \\ i=1,\cdots,k}} x^H A x = \lambda_{k+1}. \tag{1.8.14}$$

8.2 Hermite 阵的和的特征值

设 A、B、C 均为 Hermite 阵，α_i、β_i、γ_i 分别是它们的按非增次序排列的特征值，设

$$C = A + B,$$

由于

$$\gamma_k = \min_{P_i} \max_{\substack{\|x\|=1\\ P_i^H x=0,\ P_i\neq 0\\ i=1,\cdots,k-1}} (x^H C x), \tag{1.8.15}$$

仍设 $U^H A U = \mathrm{diag}(\alpha_i)$，且取 P_i 属于 U 的前 $k-1$ 列，即取$P_i = Ue_i$。记 $x = Uy$，由（1.8.15）

$$O = P_i^H x = e_i^T U^H U y = e_i^T y,$$

$$x^H A x = y^H \mathrm{diag}(\alpha_i)\, y = \sum_{i=k}^{n} \alpha_i |\eta_i|^2 \leqslant \alpha_k,$$

故

$$\gamma_k \leqslant \max_{\substack{\|x\|=1\\ P_i^H x=0,\ P_i\neq 0\\ i=1,\cdots,k-1}} (x^H C x) \leqslant \alpha_k + P_1. \tag{1.8.16}$$

由于 $A = C + (-B)$，视 C、$-B$、A 为原来的 A、B、C，类比(1.8.16)可得

$$\alpha_k \leqslant \gamma_k - \beta_n, \tag{1.8.17}$$

故

$$\alpha_k - \beta_n \leqslant \gamma_k \leqslant \alpha_k + \beta_1. \tag{1.8.18}$$

当 B 的元素满足$|b_{ij}| \leqslant \varepsilon$ 时，则由

$$-n\varepsilon \leqslant \beta_n < \beta_1 \leqslant n\varepsilon$$

可得

$$|\gamma_i - \alpha_i| \leqslant n\varepsilon, \quad i = 1, \cdots, n. \tag{1.8.19}$$

利用极大极小原理还可进一步证明

$$\gamma_{r+s-1} \leqslant \alpha_r + \beta_s, \quad \gamma + s - 1 \leqslant n. \tag{1.8.20}$$

下面我们还将介绍一个有关 n 阶实对称矩阵和它的$n-1$阶首主子阵的特征值的定理。

定理1.8.1　设A为n阶实对称阵，A_{n-1}是它的$n-1$阶首主子阵，若以$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$，$\lambda'_1 \geqslant \lambda'_2 \geqslant \cdots \geqslant \lambda'_{n-1}$分别表示它们的特征值，则诸$\lambda_i$为诸$\lambda'_i$在弱的意义下所隔离。

证明　记$I' = \mathrm{diag}(1, \cdots, 1, 0)$，

$$\boldsymbol{x}^T = (\xi_1, \cdots, \xi_n) = \boldsymbol{x}'^T + (0, \cdots, 0, \xi_n),$$

$$\boldsymbol{A} = \boldsymbol{A}' + \begin{pmatrix} 0 & \vdots & a \\ \cdots & \cdots & \cdots \\ a^T & \vdots & a_{nn} \end{pmatrix}.$$

考察

$$\min_{P_i} \max_{\substack{\|x'\|=0 \\ P_i^H x'=0,\, P_i \neq 0 \\ i=1,\cdots,s-1}} \boldsymbol{x}'^H A \boldsymbol{x}' = \min_{P_i} \max_{\substack{\|x'\|=1 \\ P_i^H x=0,\, P_i \neq 0}} \boldsymbol{x}'^H A' \boldsymbol{x}' = \lambda'_s, \tag{1.8.21}$$

而左端又可写为

$$\min_{P_i} \max_{\substack{\|x\|=1 \\ P_j^H x=0,\, P_j \neq 0 \\ j=1,\cdots,s-1 \\ e_n^T x=0}} \boldsymbol{x}^H A \boldsymbol{x} \geqslant \min_{P_i} \max_{\substack{\|x\|=1 \\ P_i^H x=0,\, P_i \neq 0 \\ i=1,\cdots,s}} (\boldsymbol{x}^H A \boldsymbol{x}) = \lambda_{s+1},$$

故

$$\lambda'_s \geqslant \lambda_{s+1}.$$

在（1.8.21）右端中，当换x'为x时，其值应是非减的，即

$$\lambda_s \geqslant \lambda'_s.$$

最后，有

$$\lambda_{s+1} \leqslant \lambda'_s \leqslant \lambda_s, \tag{1.8.22}$$

即λ_i被λ'_i在弱的意义下所隔离。

第一章 习 题

1.1 试将矩阵

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 5 \\ 3 & 4 & 7 \end{pmatrix},\quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},\quad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

分解成等秩的列矩阵和行矩阵的乘积.

1.2 设$A\in \boldsymbol{C}^{n\times m}$且 A 是秩 1 的，试证存在非零的向量 $\boldsymbol{u}\in \boldsymbol{C}^n$，$\boldsymbol{v}\in \boldsymbol{C}^m$ 使 $A=\boldsymbol{u}\boldsymbol{v}^T$，此种 $\boldsymbol{u}$、$\boldsymbol{v}$ 唯一否?

1.3 求矩阵

$$A=\begin{pmatrix} 7 & 1 & & \\ & 7 & 0 & \\ & & 7 & 1 \\ & & & 7 \end{pmatrix}$$

的线性独立的各级根向量.矩阵 A 是亏损阵吗? 是减次的吗?

1.4 设

$$A=\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\alpha_4 & -\alpha_3 & -\alpha_2 & -\alpha_1 \end{pmatrix},$$

问A可否有形如 $\boldsymbol{x}^T=(0, \times, \times, \times)$的特征向量? 如具有形如 $\boldsymbol{x}^T=(1, \times, \times, \times)$的特征向量，试求之.给出 A 非亏损的一个充分条件.

1.5 通过对行列式的直接展开，证明 $E=I-\sigma\boldsymbol{u}\boldsymbol{v}^H$的 $\det(E)=1-\sigma\boldsymbol{v}^H\boldsymbol{u}$.

1.6 用 Gerschgorin 定理说明矩阵

$$\begin{pmatrix} 9 & 1 & -2 & 1 \\ 0 & 8 & 1 & 1 \\ -1 & 0 & 7 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

至少有二实根。

1.7 设 $B\in C^{n\times n}$ 适合 $|b_{ii}|>\sum_{\substack{j=1\\j\neq i}}^{n}|b_{ij}|$，$(i=1,\cdots,n)$，

试证 $\det(B)\neq 0$。

1.8 根据特征值的隔离定理说明

$$A=\begin{pmatrix}1&2&3\\2&4&5\\3&5&6\end{pmatrix}$$

的特征值的分布情况。

1.9 设 H 是实对称阵，它在正交相似变换下化为

$$R^THR=\begin{pmatrix}1.02&-0.3&0.1\\-0.3&0.87&-0.1\\0.1&-0.1&7.23\end{pmatrix},$$

试估计 H 的特征值。

1.10 就矩阵

$$H=\begin{bmatrix}4&3i\\-3i&2\end{bmatrix}$$

验证极值原理。

1.11 设 $A\in C^{n\times n}$ 且 A 的各列是互相正交的单位向量，证明 $\max\limits_{\|x\|=1}|(Au,u)|$ 只能在 u 是对应于 A 的依模最大的特征值的特征向量时取得。

1.12 证明任意方阵可经由酉相似变换化为上三角阵。

1.13 证明同时是酉阵又是三角阵的矩阵必是对角阵，且对角元依模等于1，证明同时是正规阵又是三角阵的矩阵必定是对角阵。

1.14 设 $F\in C^{n\times m}(n\geqslant m)$，$F$ 的秩为 m，证明存在唯一的 $P_F\in C^{n\times n}$，使对任意的 $x\in R^n$ 有

$$x=P_Fx+(x-P_Fx),$$

其中 P_F 是 F 的列的组合，$(x-P_Fx)$ 与 F 的任意的列正交，试导出 P_F 的表达式。

1.15 设 $T=(t_{ij})$ 是上三角阵，当 $t_{ii}\neq t_{jj}$ 时，可以适当选择 δ 使当 $j>i$ 时，有

$$e_i E(e_i, e_j - \sigma) T E(e_i, e_j, \sigma) = 0,$$

从而证明对任意矩阵 A，存在矩阵 $X = \Omega$，其中 ΩF 是初等 Hermite 阵的乘积，F 是形如 $E(e_i, e_j, \sigma)$ 的初等矩阵的乘积，使得

$$X^{-1} A X = \mathrm{diag}(R_1, R_2, \cdots),$$

其中每个 R_i 是三角阵且其对角元是彼此相等的。

1.16 幂零矩阵只能有零特征值，幂等矩阵只能有 0 和 1 的特征值。此处幂零和幂等矩阵分别是指满足方程 $A^\nu = 0 \neq A^{\nu-1}$ 和 $A^2 = A$ 的矩阵，这里 ν 是某一正整数，称为 A 的幂零指数。

1.17 设 A 的幂零指数是 ν，则存在向量 v_1，其Krylov序列 $v_1, v_2, \cdots, v_k \cdots$ 定义为

$$v_{k+1} = A v_k \, (k = 1, 2, \cdots)$$

满足 $v_{\nu+1} = 0 \neq v_\nu$，记 $V = (v_1, \cdots, v_\nu)$ 则有

$$AV = VJ,$$

1.18 设 A、B 均是可正规化矩阵则下之二关系等阶：

(1) $AB = BA$,

(2) 存在非奇异阵 X 使 $X^{-1}AX, X^{-1}BX$ 同是对角阵。

1.19 用两种方法展开 $\det\begin{pmatrix} \lambda I & -A \\ -B & \lambda I \end{pmatrix}$ 证明 AB，BA 的特征多项式除一个 λ 幂因子外是相等的。

1.20 设 A、B 均为方阵，证明存在 $X \neq 0$ 满足 $AX = XB$ 的充要条件是 A、B 至少有一个公共的特征值，据此证明 $AX - XB = C \neq 0$ 有唯一解的充要条件是 A、B 没有公共特征值。

1.21 如果 A^ν 的迹数对一切正整数 $\nu > 0$ 均为零，则 A 必为幂零矩阵，反之也成立。

1.22 设 $A = J + \varepsilon^n e_1 e_n^T$，其中

$$J = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & \ddots & & \\ & & \ddots & & 0 \\ & & & 1 & 0 \end{pmatrix},$$

证明 A 的特征值为 $\omega^k\varepsilon$，这里 $\omega=\exp(i\pi/n)$。

1.23　设 $F=J-fe_n^T$，$f^T=(f_1,\cdots,f_n)$，验证

$$\det(F-\lambda I)=\begin{vmatrix} -\lambda & -\lambda^2 & \cdots & \lambda^{n-1} & -\varphi(\lambda) \\ 1 & 0 & & & \\ & 1 & \ddots & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{vmatrix},$$

这里 $\varphi(\lambda)=\lambda^n+f_n\lambda^{n-1}+\cdots+\lambda f_2+f_1$。

1.24　与对角阵可换的矩阵的一般形式是什么与 Jordan 块可换的矩阵的一般形式又是什么？

1.25　任意实矩阵 A 可唯一地分解成一个实对称阵 B 和一个实反对称阵 C 之和，而 A 为实正规阵的充要条件是 B、C 可换。

参　考　书

1. J. H. Wilkinson *"The Algebraic Eigenvalue Problems"*. England. Oxford Univ. Press.1965.
2. Alston S. Householder *"The Theory of Matrices in Numerical Analysis"*. Blaisdell, New York. 1964.
3. Joel N. Franklin. *"Matrix Theory"*. Prentice, Inc., Englewood Cliffs, *New Jorsey*.

第二章　误差分析

数值代数中提供的数值解法，通常需要在高速电子计算机上进行。代数四则运算通常都带有舍入误差，原始数据也可能有误差，因此，求得的代数问题的解往往是近似解，近似解的精度不但与问题的性态有关，而且与算法也有关。本章将利用“向后误差分析”的方法来讨论方程组的性态问题和算法稳定性等有关概念。

§1　方程组的性态和算法的稳定性

1.1 方程组的性态

用数值方法解具体问题时，我们最关心的问题是解的精度如何？通过实际计算，我们发现，解的精度与原始数据的精度，数学模型和算法等都有关系。下面从数值问题的一般概念出发，逐步介绍与数值问题的解的精度有关的内容。

例2.1.1　线性方程组

$$\begin{bmatrix} 1 & 10^4 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 10^4 \\ 2 \end{bmatrix}$$

的精确解如按三位有效数字进行舍入为：$x_1=x_2=1.00$。如果对原始数据和中间结果均取三位有效数字利用部分主元素法求解，得到

$$x_1=0.00,\ x_2=1.00;$$

利用全主元法求解，得到

$$x_1=1.00,\ x_2=1.00.$$

例2.1.2 矩阵

$$A=\begin{bmatrix}1 & 10^4\\ 1 & 1\end{bmatrix}$$

如果对元素取三位有效数字，利用特征多项式求其特征值，得到

$$\lambda_1=101,\ \lambda_2=-99.$$

一般地，我们可以把数值问题描述成从原始数据 $d\in \mathbf{D}$ 到数值结果 $s\in S$ 的映射 ϕ，即

$$\phi:\ D\to S, \tag{2.1.1}$$

其中 D 表示原始数据 d 的集合，S 表示数值结果 S 的集合。这里，ϕ 包括数值问题的描述和解法等双重意义，ϕ 的自变量 d 是所论问题的原始输入数据，而 $\phi(d)$（即 S）就是问题的解。在例 2.1.1 中

$$d=\begin{bmatrix}1 & 10^4 & 10^4\\ 1 & 1 & 2\end{bmatrix}.$$

我们可分别用 $S=\phi(d)$，$S_1=\phi_1(d)$，$S_2=\phi_2(d)$ 表示相应问题的方程组的精确解，用部分主元素法求得的解和用全主元法求得的解。此时

$$S=\phi(d)=\begin{bmatrix}\dfrac{10000}{9999}\\ \dfrac{9999}{9999}\end{bmatrix},\quad S_1=\phi_1(d)=\begin{bmatrix}0.00\\ 1.00\end{bmatrix},$$

$$S_2=\phi_2(d)=\begin{bmatrix}1.00\\ 1.00\end{bmatrix}.$$

将某个问题化成数值问题时，我们往往希望，所求的解 $\phi(d)$ 存在而且唯一。如果用非负数 $\|\phi(d+\delta d)-\phi(d)\|_s$ 来表示用两组数据 d 和 $d+sd$ 所求的相应解的差异，当

(1) 存在 $\varepsilon>0$，使对所有满足 $\|\delta d\|_D<\varepsilon$ 的 δd，解 $\phi(d+\delta d)$ 均存在而且唯一；

(2) 当 $\|\delta d\|_D \to 0$ 时，$\|\phi(d+\delta d)-\phi(d)\|_S \to 0$，则称该问题对给定的数据 d 是适定的。其中 $\|\cdot\|_D$，$\|\cdot\|_S$ 分别表示定义在 D 和 S 上的某种范数。条件（2）表示解对原始数据的连续依赖性。

如果问题的解多于一个，或者解不连续依赖于原始数据，那么，应用数值方法求解时，就会产生求得的解是否可信的问题，因此，本书只限于讨论适定的问题。

在实际问题中，我们往往只知道原始数据的近似值 $\bar{d}$，对于确定的 ϕ，一般只能计算 $\phi(\bar{d})$。如果 $\|d-\bar{d}\|_D \leqslant \delta$，而 $\|\phi(\bar{d})-\phi(d)\|_S \leqslant \varepsilon$，其中 ε 可以估计成较小的正数，我们就称该计算问题是"良态"的。然而，如 $\|\phi(\bar{d})-\phi(d)\|$ 很大，我们就称该计算问题是"病态"的。对于病态问题，当原始数据有微小的变化时，可能引起数值结果有很大的波动。这时若要求数值结果达到某种精度，就必须要求原始数据具有更高的精度才行。有时，对于固定的 ϕ 和某种运算工具（如电子计算机），要求数值结果达到指定的精度甚至是不可能的。值得注意的是：对于同一组原始数据 d，关于不同的 ϕ，其性态也可能不同。

例2.1.3　线性方程组

$$\begin{bmatrix} 3.000 & 4.127 \\ 1.000 & 1.374 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 15.41 \\ 5.147 \end{bmatrix} \tag{2.1.2}$$

有解　$x_1 = 13.6658$，$x_2 = -6.2$

但线性方程组 $$\begin{bmatrix} 3.000 & 4.122 \\ 1.000 & 1.374 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 15.41 \\ 5.147 \end{bmatrix} \tag{2.1.3}$$

没有解。方程组(2.1.2)与(2.1.3)仅在系数矩阵的元素(1,2)上有0.005的差异，但前者有解，后者却没有解。这是由于方程组(2.1.2)是"病态"所引起的。方程组的病态也可以说成系数矩阵关于方程组求解问题是"病态"的。

例2.1.4　矩阵

$$A=\begin{bmatrix}3.000 & 4.127\\1.000 & 1.374\end{bmatrix}$$

的特征值是$\lambda_1=-0.001$，$\lambda_2=4.375$，而矩阵

$$B=\begin{bmatrix}3.000 & 4.122\\1.000 & 1.374\end{bmatrix}$$

的特征值为$\lambda_1=0$，$\lambda_2=4.374$。

在例2.1.4中，矩阵A与B在元素(1,2)上同样也有0.005的差异，但是它们所求得的特征值仅有0.001的差别，这是由于矩阵A的特征值问题是“良态”的。这两个例子说明了问题的性态与问题的描述有关。关于方程组的性态问题，在国内外已有很多研究，近年来，随着电子计算机的计算速度和存贮量的提高，不少原来计算中遇到的难题已得到了解决。因此，愈来愈多的人对研究“病态”问题发生了兴趣，并且取得了不少研究成果。

本节着重研究线性方程组的性态，特别是要讨论如何从数值上去判定线性方程组的性态等问题。

首先我们来看一下方程组(2.1.2)。根据高等代数的知识可知，线性方程组的解与系数行列式密切相关。该方程组的系数行列式为

$$D=\begin{vmatrix}3.000 & 4.127\\1.000 & 1.374\end{vmatrix}=-0.006,$$

其绝对值较小。那么，是否系数行列式的绝对值小，方程组就是病态的呢？下面先看两个例题。

例2.1.5 线性方程组

$$\begin{cases}0.221x_1+0.054x_2=0.329\\0.049x_1+0.034x_2=0.117\end{cases}\qquad(2.1.4)$$

的系数行列式为

$$D=\begin{vmatrix}0.221 & 0.054\\0.049 & 0.034\end{vmatrix}=0.004868,$$

其解为

$$x_1=1,\ x_2=2.$$

线性方程组

$$\begin{cases}0.221x_1+0.059x_2=0.329;\\0.049x_1+0.034x_2=0.117\end{cases}$$

的解为

$$x_1=0.93,\ x_2=2.1$$

从例 2.1.5 可以看出，系数行列式的绝对值小，方程组并不一定病态。那么，用什么数值来“度量”方程组的性态才合理呢？方程组(2.1.2)与(2.1.4)，虽然系数行列式的绝对值都较小，但是两者却有很大不同，前者的系数矩阵的列几乎线性相关，而后者却没有这一特点。方程组(2.1.2)可以改写成

$$\begin{pmatrix}3.000\\1.000\end{pmatrix}x_1+\begin{pmatrix}4.127\\1.374\end{pmatrix}x_2=\begin{pmatrix}15.41\\5.147\end{pmatrix}.$$

把 x_1 和 x_2 的系数及常数项记作向量 $\boldsymbol{a}_1$，$\boldsymbol{a}_2$，$\boldsymbol{b}$，则方程组(2.1.2)可以写成

$$\boldsymbol{a}_1x_1+\boldsymbol{a}_2x_2=\boldsymbol{b}.$$

今把向量 $\boldsymbol{a}_1$，$\boldsymbol{a}_2$，$\boldsymbol{b}$ 标准化，记 $\overline{\boldsymbol{a}}_1=\dfrac{\boldsymbol{a}_1}{\|\boldsymbol{a}_1\|}$，$\overline{\boldsymbol{a}}_2=\dfrac{\boldsymbol{a}_2}{\|\boldsymbol{a}_2\|}$，$\overline{\boldsymbol{b}}=\dfrac{\boldsymbol{b}}{\|\boldsymbol{b}\|}$（则 $\|\overline{\boldsymbol{a}}_1\|=\|\overline{\boldsymbol{a}}_2\|=\|\overline{\boldsymbol{b}}\|=1$），组成新方程组

$$\overline{\boldsymbol{a}}_1y_1+\overline{\boldsymbol{a}}_2y_2=\overline{\boldsymbol{b}},\qquad(2.1.5)$$

其中，

$$\frac{\|\boldsymbol{a}_1\|}{\|\boldsymbol{b}\|}x_1=y_1,\quad \frac{\|\boldsymbol{a}_2\|}{\|\boldsymbol{b}\|}x_2=y_2.$$

倘若(2.1.5)的解y_1，y_2已经求出，并把关系式

$$\overline{\boldsymbol{b}}-\overline{\boldsymbol{a}_2}y_2=\overline{\boldsymbol{a}_1}y_1$$

画在平面坐标系$O\text{-}xy$上，如图2.1。

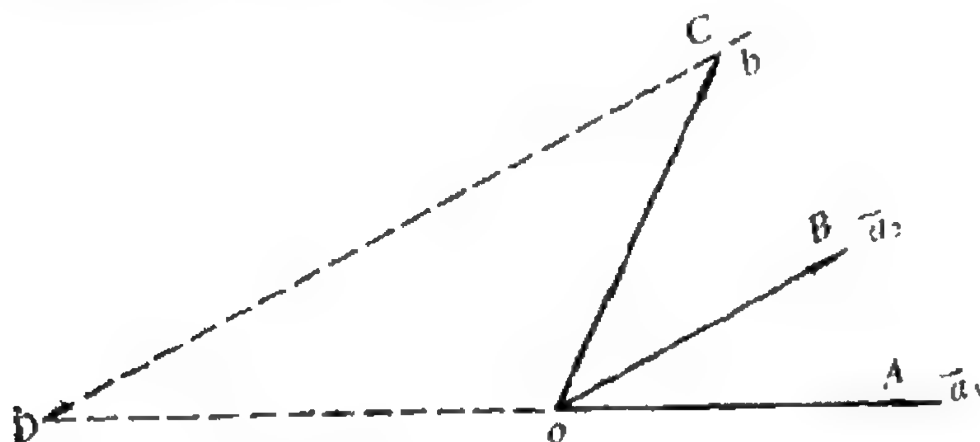


图2.1 列向量的变化对解的影响

其中$CD//\overline{\boldsymbol{a}_2}$，$OD$是$AO$的延长线，故有

$$\overline{\boldsymbol{b}}-\overline{\boldsymbol{a}_2}y_2=\overline{\boldsymbol{a}_1}y_1,$$

从而方程(2.1.5)的解的绝对值$|y_1|$和$|y_2|$分别是线段DO和CD的长。从图2.1中可以清楚地看到，当$\overline{\boldsymbol{a}_1}$和$\overline{\boldsymbol{a}_2}$几乎线性相关时，即向量$\overline{\boldsymbol{a}_1}$和$\overline{\boldsymbol{a}_2}$的夹角很小时，$\overline{\boldsymbol{a}_1}$和$\overline{\boldsymbol{a}_2}$稍有变动，则向量$DO$和$CD$的长就可能发生很大的变化，因而方程组是病态的。

上面我们从直观上说明了系数矩阵的列向量的线性关系与方程组的性态有着密切的关系。把图形中的线性关系变成数值关系，从而去判定方程组的性态，将是很必要的。

设线性方程组为

$$A\boldsymbol{x}=\boldsymbol{b},$$

其中

$A=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n)$，$\boldsymbol{a}_i$是$A$的第$i$个列向量，$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^T$，$\boldsymbol{b}=(\alpha_1,\alpha_2,\cdots,\alpha_n)^T$。根据线性代数知识，列向量系$\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_n\}$为线性相关的充要条件是，该向量系的Gram行列式

$G[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n] = \det(\boldsymbol{\alpha}_i^T \cdot \boldsymbol{\alpha}_j) = 0$。Gram 行列式的几何意义是：以向量系$\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n\}$中各列向量为边所组成的平行多面体的体积的平方，若向量系线性相关，则该体积为零。显然，向量系中的非零向量是否线性相关与它们本身的长度是无关的，但是体积的大小却与向量的长度有关。为了使$G[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \cdots, \boldsymbol{\alpha}_n]$能表明向量系的相关程度，以后，我们可以将向量系中的向量标准化，即使$\|\boldsymbol{\alpha}_i\| = 1\ (i = 1, 2, \cdots, n)$。若

$$G[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n] \leqslant \varepsilon,$$

则称向量系$\{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n\}$为**$\varepsilon$-线性相关。**

据第一章§3的知识，对于实非奇异矩阵$\boldsymbol{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n)$必存在正交矩阵$\boldsymbol{Q}$使

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R},$$

其中

$$\boldsymbol{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & r_{nn} \end{pmatrix},$$

$r_{ii} > 0\ \ (i = 1, 2, \cdots, n)$，而且分解是唯一的。因此

$$\boldsymbol{A}^T\boldsymbol{A} = (\boldsymbol{Q}\boldsymbol{R})^T(\boldsymbol{Q}\boldsymbol{R}) = \boldsymbol{R}^T\boldsymbol{R},$$

从而

$$\begin{aligned} G[\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_n] &= \det(\boldsymbol{A}^T\boldsymbol{A}) = (\det(\boldsymbol{R}))^2 \\ &= r_{11}^2 \cdot r_{22}^2 \cdot \cdots \cdot r_{nn}^2. \end{aligned}$$

由于列向量$\boldsymbol{\alpha}_i$均已单位化，所以$|r_{ii}| \leqslant 1$，$i = 1, 2, \cdots, n$；设

$$p_1 = \mathrm{Max}(|r_{11}|, |r_{22}|, \cdots, |r_{nn}|);$$

$$p_n = \mathrm{Min}(|r_{11}|, |r_{22}|, \cdots, |r_{nn}|).$$

因此，p_n愈小，向量系愈接近相关。

定义2.1.1 称 $k^{(1)}(A)=\frac{1}{p_n}$ 为矩阵**A**的一阶**条件数**。当 p_n 很小或 $k^{(1)}(A)$ 很大时，向量系就接近线性相关，并称方程组是病态的。

定义2.1.1中所述矩阵**A**的一阶条件数 $k^{(1)}(A)$ 与数值分析中定义的矩阵**A**的条件数 $\mathrm{Cond}(A)=\|A\|\cdot\|A^{-1}\|$ 虽然形式不同，但它们却是从不同角度刻划了线性方程组的性态，在表达形式和使用方面也各具特色。它们之间还有如下关系：

定理2.1.1 设 $k^{(1)}(A)$ 是矩阵**A**的一阶条件数，$\mathrm{Cond}(A)=\|A\|\|A^{-1}\|$ 是通常的条件数，则

$$K^{(1)}(A)\leqslant \mathrm{Cond}(A)\leqslant\sqrt{n}\cdot 2^{n-1}K^{(1)}(A). \tag{2.1.6}$$

这个定理的证明，可以参看高等学校计算数学学报1979年第一期何旭初先生的文章。

为了说明 $K^{(1)}(A)$ 的大小能刻划方程组的性态，只需说明初始数据有微小变化时，$K^{(1)}(A)$ 的大小对方程组的解有多大影响就行了。根据定理2.1.1，只需对 $\mathrm{Cond}(A)$ 进行这样的说明即可。为此，设方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 中，**A**和 $\boldsymbol{b}$ 分别有误差 δA 和 $\delta\boldsymbol{b}$，从而解有误差 $\delta\boldsymbol{x}$，则它们之间有如下关系：

定理2.1.2 设在 $A\boldsymbol{x}=\boldsymbol{b}\neq 0$ 中，**A**是非奇异阵，且

$$(A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}.$$

如果 $\|A^{-1}\|\|\delta A\|<1$，则

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\mathrm{Cond}(A)}{1-r\mathrm{Cond}(A)}\left(r+\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right), \tag{2.1.7}$$

其中

$$r=\frac{\|\delta A\|}{\|A\|}.$$

证明 由 $(A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}$

可得

$$(A+\delta A)\delta \boldsymbol{x}=\delta \boldsymbol{b}-(\delta A)\boldsymbol{x},$$

从而有

$$A(I+A^{-1}\delta A)\delta \boldsymbol{x}=\delta \boldsymbol{b}-(\delta A)\boldsymbol{x}. \qquad (2.1.8)$$

因为

$$\|A^{-1}\delta A\|\leqslant\|A^{-1}\|\|\delta A\|<1,$$

所以$(I+A^{-1}\delta A)^{-1}$存在，于是由(2.1.8)得到

$$\delta \boldsymbol{x}=(I+A^{-1}\delta A)^{-1}A^{-1}(\delta \boldsymbol{b}-\delta A\cdot \boldsymbol{x}),$$

故

$$\|\delta \boldsymbol{x}\|\leqslant\frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|}(\|\delta \boldsymbol{b}\|+\|\delta A\|\|\boldsymbol{x}\|),$$

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{x}\|}+\|\delta A\|\right).$$

又因为

$$\|A\|\|\boldsymbol{x}\|\geqslant\|\boldsymbol{b}\|,$$

所以

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|\delta A\|}\left(\frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|}+\frac{\|\delta A\|}{\|A\|}\right),$$

即

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{\mathrm{Cond}(A)}{1-r\cdot\mathrm{Cond}(A)}\left(r+\frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|}\right).$$

推论2.1.1 在定理2.1.2中，如果$\delta \boldsymbol{b}=0$，则

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\frac{r\cdot\mathrm{Cond}(A)}{1-r\cdot\mathrm{Cond}(A)}; \qquad (2.1.9)$$

如果$\delta A=0$，则

$$\frac{\|\delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\mathrm{Cond}(A)\frac{\|\delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|}. \qquad (2.1.10)$$

从(2.1.7)，(2.1.9)，(2.1.10)可以看到，如果 cond(A) 很大，那末，解的相对误差界将比$\|\delta A\|/\|A\|$和$\|\delta \boldsymbol{b}\|/\|\boldsymbol{b}\|$大很多，有可能使方程组的计算解与精确解相差很大，即原始数据有较小的变化时，方程组的解可能有很大的变化，也就是方程组可能是病态的。

关于矩阵性态的定量的研究，还在继续发展，怎样用便于估计的量去全面刻划线性方程组的性态问题，引起了不少人的兴趣。目前发表的论条件数的各种定义，正是反映了这方面的努力。

1.2 算法的稳定性概念

数学问题在求解过程中，往往需要进行成千上万次的运算，由于舍入的原因，每一步运算都会带来误差，这种舍入误差的积累是否会影响解的精度呢？这就要看算法的"好坏"了。在数学上，衡量算法"好坏"的主要标志之一是看算法是否稳定。从输入数据中所含的误差对解的精度的影响以及从输出结果来推测输入数据中的误差来分析，有两种稳定性的定义，即向前稳定与向后稳定的定义值得提出来。

"向前稳定"是从"向前误差分析"所得到的结果来衡量的。所谓向前误差分析，是指对每一步计算，都去找出精确值与计算值之间的误差界（此时，输入数据的误差界被认为是设定的），随着计算过程的向前推移而逐步向前分析，直到得出最后计算结果与精确结果之间的误差界为止。如果这个界限比较小，或者在所"允许"的范围之内，我们就称这个算法是稳定的，更确切地说，就叫做"向前稳定"。

向前误差分析在使用时，由于精确值事先不知道，每步计算都要估计误差限，不但工作量大，而且容易产生过估现象。因此，使用向前误差分析去分析算法的稳定性时，历史上曾出

现过对某些算法的悲观论调。在很长一段时间内，研究误差分析的进展不大。直到提出“向后误差分析”之后，误差的分析研究才获得了比较全面的进展。

“向后稳定”是由向后误差分析的结果来判定的。设以 d 表示某数学问题的输入数据，$d \in \mathbf{D}$，用 $\phi(d)$ 表示该问题的精确解，即求解的每一步中的中间结果都是精确地进行的，不带舍入误差的，而以 $\phi^*(d)$ 表示同一问题的计算解，即求解的每一步都是考虑了舍入误差的。如果把 $\phi^*(d)$ 写成 $\phi^*(d) = \phi(\bar{d})$ 的形式，$\bar{d} \in \mathbf{D}$，并且 d 与 $\bar{d}$ 的差别即 $\|d - \bar{d}\|$ 也不太大时，我们就称算法 ϕ^* 关于 $\mathbf{D}$ 是稳定的。可以把 $\bar{d}$ 看成是 d 的带有误差的输入数据。由于原始数据一般总有误差，因此，这样处理是合于实际的。把计算过程的误差返回到原始数据的误差，这样的误差分析方法通常称为向后误差分析方法，如果算法经这样的分析是稳定的，则称之为是向后稳定的。

用一个稳定算法去解良态问题时，由于计算解 $\phi^*(d)$ 必须接近某个精确解 $\phi(\bar{d})$，其中 $\bar{d}$ 接近于 d。又由于问题是良态的，即当 $\bar{d}$ 接近 d 时，$\phi(\bar{d})$ 接近 $\phi(d)$，因此 $\phi^*(d)$ 也接近于 $\phi(d)$。也就是说，用一个稳定算法去解良态问题时，计算解必定接近于精确解。如果用一个稳定算法去解病态问题，那末，由于 $\bar{d}$ 接近于 d 时，$\phi(\bar{d})$ 不一定接近于 $\phi(d)$，从而 $\phi^*(d)$ 不一定接近于 $\phi(d)$，也就是说，稳定算法解病态问题时，得到的计算解与精确解可能差别很大。对于病态问题的求解，不是一般稳定算法所能解决的，必须要用特殊的算法来处理。

§2 误差分析

前一节我们讨论了数学问题的性态和算法的稳定性，这两个问题都是由于数据或运算中出现误差而引起的。为了深入地

讨论这些问题，必须进行误差分析。在四十年代以前，几乎没有人系统地研究过这个问题。直到四十年代中期，由于电子计算机的出现，才引起人们的巨大兴趣，开始的一些分析，然而，结论是很悲观的。例如，认为消去法不能用于解高阶方程组问题等。到1947年 Von Neumann 等的文章才澄清了这种悲观的论调，并含蓄地引进了“向后误差分析”的思想。直到1954年，Givens的文章才明确了“向后误差分析”的方法。在 J. H. Wilkinson 写的“The Algebraic Eigenvalue Problem”一书中，应用这种方法系统地分析了数值代数中众多的具体的算法，得到相当满意的结果。限于篇幅，本节将主要使用向后误差分析的方法，对代数中的基本运算和消去法进行浮点误差分析。

2.1 浮点基本运算的误差分析

对于浮点计算的各种模式，我们不能一一介绍，这里仅采用一种规格化的浮点计数系统来说明舍入误差分析方法。

对于任何实数x，通常都可找到一个整数 a 及一个小数b，使

$$x = 2^a \cdot b. \tag{2.2.1}$$

为了使这种表示形式是唯一确定的，可以规定

$$x \neq 0\text{时，} \frac{1}{2} \leqslant |b| < 1, \tag{2.2.2}$$

$$x = 0\text{时，} a = b = 0. \tag{2.2.3}$$

数 b 通常被表示成二进制形式，它可以是有限位的也可以是无限位的。在电子计算机上通常用若干位二进制数位，例如用 t 位二进制数位表示 b，当 b 的二进制数位多于 t 位时，就要在 b 的第$(t+1)$位上进行舍入，即当 b 的第 $(t+1)$位为 1 时，就在 b 的第 t 位上进 1，同时把 b 的从$(t+1)$位以后的数甩掉，而

当 b 的第$(t+1)$位是零时，就只甩掉$(t+1)$位以后的数，而不再在 t 位上进位。这样得出的数的表示形式，叫做 x 的浮点规格化的机器表示形式。简称为机器表示形式。a 叫做 x 的阶码，b 叫做尾数。由此可见数 x 与其机器表示的值可能并不精确地相等，如果有误差，则造成误差的原因是“舍入”，而误差的大小取决于 t ——机器的字长。即

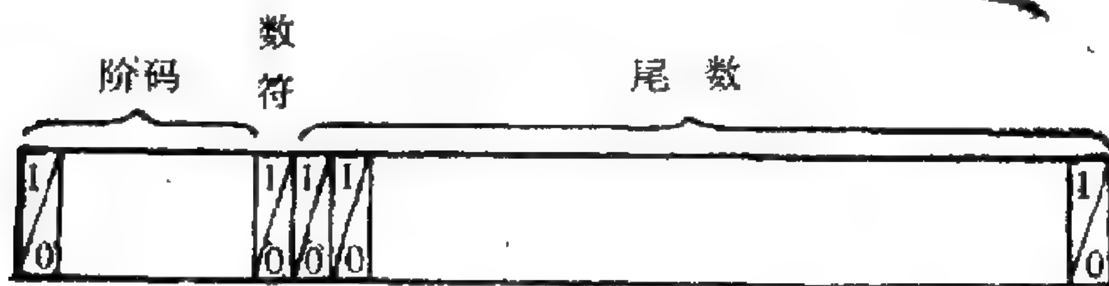


图2.2 数的机器表示

如果用$fl(x)$来表示 x 的机器表示值，则含于 $fl(x)$ 中的绝对误差 ε_e 和相对误差 ε_r 应满足

$$fl(x)=x+\varepsilon_e,\quad |\varepsilon_e|\leqslant 2^{a-t-1},$$

$$fl(x)=x+x\cdot\frac{\varepsilon_e}{x}=x\left(1+\frac{\varepsilon_e}{x}\right)=x(1+\varepsilon_r),$$

$$|\varepsilon_r|=\left|\frac{\varepsilon_e}{x}\right|\leqslant\frac{2^{a-t-1}}{|2^a\cdot b|}\leqslant 2^{-t}.$$

下面我们讨论算术运算的舍入误差，由于不可能同时研究所有的计算机，在此假定计算机的累加器是双字长的，即可以接受双字长的数。

设 x，y 为规格化的浮点数，并且用

$$fl(x+y),\ fl(x-y),\ fl(x\times y),\ fl(x/y)$$

分别表示进行浮点加，减，乘，除四则运算后的结果。

2.1.1 加法(减法)

设 $x=2^{b_1}a_1$，$y=2^{b_2}a_2$，$b_1>b_2$；求 $fl(x\pm y)$。

如果$b_1-b_2\leqslant t$，则将a_2右移b_1-b_2位（将移出去的b_1-b_2位保留在累加器的后半部），使x_1与x_2同阶，然后求$a_1\pm a_2\times 2^{b_2-b_1}$，并进行规格化，对前$t$位进行舍入（即在$t+1$位上加上"±1"），最后再调整阶码。其中只有一步舍入。设最后结果为$2^{b_3}a_3$，则

$$fl(x\pm y)=x\pm y+\varepsilon_e,\quad |\varepsilon_e|\leqslant 2^{b_3-t-1},$$

$$fl(x\pm y)=(x\pm y)+(x\pm y)\frac{\varepsilon_e}{x\pm y}$$

$$=(x\pm y)\left(1+\frac{\varepsilon_e}{x\pm y}\right)=(x\pm y)(1+\varepsilon_r),$$

$$|\varepsilon_r|=\frac{|\varepsilon_e|}{|x\pm y|}\leqslant\frac{2^{b_3-t-1}}{|2^{b_3}a_3|}\leqslant\frac{2^{b_3-t-1}}{2^{b_3}-1}=2^{-t}.$$

于是可以写成如下形式

$$\begin{cases}fl(x\pm y)=(x\pm y)(1+\varepsilon_r),\\ |\varepsilon_r|\leqslant 2^{-t}\qquad (b_1\geqslant b_2).\end{cases}\tag{2.2.4}$$

当$b_1\leqslant b_2$时，(2.2.4)同样满足。(2.2.4)表明x与y的计算和(或差)是$x(1+\varepsilon_r)$与$y(1+\varepsilon_r)$的精确和（或差），这里$|\varepsilon_r|\leqslant 2^{-t}$。

2.1.2 乘法

设 $x=2^{b_1}\cdot a_1$，$y=2^{b_2}\cdot a_2$ 求$fl(x\times y)$。

首先是指数相加：b_1+b_2，然后精确求出$a_1\cdot a_2$的双字长，即$2t$位的结果，再规格化，调整阶码，最后将$2t$位字长的数舍入成t位。若x或y为零，则乘积直接取零。由此可见也只有最后一步舍入。设其结果为$2^{b_4}\cdot a_4$，则

$$fl(x\times y)=x\times y+\varepsilon_e,\quad |\varepsilon_e|\leqslant 2^{b_4-t-1},$$

$$fl(x\times y)=(x\times y)\left(1+\frac{\varepsilon_e}{x\times y}\right)=(x\times y)(1+\varepsilon_r),$$

$$|\varepsilon_r|=\frac{|\varepsilon_E|}{|x\times y|}\leqslant\frac{2^{b_4-t-1}}{|2^{b_4}\cdot a_4|}\leqslant 2^{-t},$$

从而有

$$\begin{cases} fl(x\times y)=(x\times y)(1+\varepsilon_r), \\ |\varepsilon_r|\leqslant 2^{-t}, \end{cases} \tag{2.2.5}$$

即 x 与 y 的计算积是 $x(1+\varepsilon_r)$ 与 y 或 x 与 $y(1+\varepsilon_r)$ 或 $x(1+\varepsilon_r)^{\frac{1}{2}}$ 与 $y(1+\varepsilon_r)^{\frac{1}{2}}$ 的精确积，其中 $|\varepsilon_r|\leqslant 2^{-t}$。这三种形式是可以视分析的方便而任选的。

例2.2.1 设 $x=2^{101}\cdot 0.101101$ $y=2^{11}\cdot 0.111101$ 求 $fl(x-y)$ 和 $fl(x\times y)$。

解 (1) 因为 $t=110$ 且 $b_1-b_2=101-11=10\leqslant 110$

所以 $\bar{a}_3=a_1-a_2\times 2^{-10}=0.011110$，

将 $\bar{a}_3=0.011110$ 规格化得到 0.111100。

对 0.111100 进行舍入后得 $a_3=0.111100$。

调整阶码算得最后结果为

$fl(x-y)=2^{100}\cdot 0.111100=(x-y)(1+\varepsilon_r)$，

其中 $|\varepsilon_r|\leqslant 2^{-110}$。

(2) 因为 $a_1\cdot a_2=0.101011001001$，

将 $a_1\cdot a_2$ 舍入成 6 位，得到 $a_4=0.101011$，

$2^{b_1+b_2}=2^{101+11}=2^{1000}=2^{b_4}$，

所以

$$\begin{cases} fl(x\times y)=2^{1000}\cdot 0.101011=(x\times y)(1+\varepsilon_r), \\ |\varepsilon_r|\leqslant 2^{-110}. \end{cases}$$

2.1.3 除法

求 $fl(x/y)$，只要 $y\neq 0$，同加法和乘法一样，可以得到

$$fl(x/y)=(x/y)(1+\varepsilon_r)$$
$$|\varepsilon_r|\leqslant 2^{-t} \tag{2.2.6}$$

即 x 与 y 的计算商是 $x(1+\varepsilon_r)$ 与 y 或 x 与 $y/(1+\varepsilon_r)$ 的精确商，

其中$|\varepsilon_r|\leqslant 2^{-t}$。

综合上面的讨论，如果用θ表示加，减，乘，除四则运算中的任何一种，则有下列统一的表示：

$$\begin{cases} fl(x\theta y) = (x\theta y)(1+\varepsilon_r), \\ |\varepsilon_r| \leqslant 2^{-t}. \end{cases} \tag{2.2.7}$$

若设

$$1+\varepsilon_r = \frac{1}{1+\varepsilon_r'}，\text{其中} |\varepsilon_r| \leqslant 2^{-t},$$

则

$$1+\varepsilon_r' = \frac{1}{1+\varepsilon_r} \approx 1-\varepsilon_r+\varepsilon_r^2-\cdots+\cdots,$$

即

$$\varepsilon_r' = -\varepsilon_r+\varepsilon_r^2-\varepsilon_r^3+\cdots-\cdots.$$

当ε_r很小时，可以忽略ε_r的高次项，于是

$$\varepsilon_r' \doteq -\varepsilon_r$$

从而(2.2.7)还可以写成如下适用的形式

$$\begin{cases} fl(x\theta y) = (x\theta y)/(1+\varepsilon_r'), \\ |\varepsilon_r'| \leqslant 2^{-t}. \end{cases} \tag{2.2.8}$$

2.1.4 连加和连乘

设$x_1, x_2, \cdots, x_n$均为标准浮点数，求$fl(x_1+x_2+\cdots+x_n)$和$fl(x_1x_2\cdots x_n)$。计算从左向右进行，每进行一次都进行舍入，然后再做下一次。

(1) 连加

设

$$\begin{cases} s_1 = x_1, \\ s_2 = fl(s_1+x_1) \equiv (s_1+x_2)(1+\varepsilon_2), \\ s_k = fl(s_{k-1}+x_k) \equiv (s_{k-1}+x_k)(1+\varepsilon_k), \\ \qquad k=3,4,\cdots,n \\ |\varepsilon_k| \leqslant 2^{-t} \quad (k=2,3,\cdots,n), \end{cases} \tag{2.2.9}$$

则有

$$
\begin{cases}
s_n = fl(x_1 + x_2 + \cdots + x_n) \equiv x_1(1+\eta_1) + x_2(1+\eta_2) \\
\qquad\qquad + \cdots + x_n(1+\eta_n), \\
(1-2^{-t})^{n+1-k} \leqslant 1+\eta_k \leqslant (1+2^{-t})^{n+1-k}, \\
\quad k = 1, 2, \cdots, n.
\end{cases}
\qquad (2.2.10)
$$

证明 将(2.2.9)中的 s_k 逐项代入，可得

$$s_n = x_1(1+\eta_1) + x_2(1+\eta_2) + \cdots + x_n(1+\eta_n).$$

其中

$$1+\eta_1 = 1+\eta_2 = (1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n),$$

$$1+\eta_k = (1+\varepsilon_k)(1+\varepsilon_{k+1})\cdots(1+\varepsilon_n),$$

$$k = 3, 4, \cdots, n.$$

从而有

$$(1-2^{-t})^{n-k+1} \leqslant 1+\eta_k \leqslant (1+2^{-t})^{n-k+1}.$$

由于$(1\pm 2^{-t})^k$的形式经常出现，所以引入简化条件。设

$$k \cdot 2^{-t} < 0.1 \qquad (2.2.11)$$

此条件并不苛刻，例如当 $t=40$ 时，k可达10^{11}。

由于

$$(1+2^{-t})^k \leqslant 1 + k\cdot 2^{-t} + \frac{k(k-1)}{2!}(2^{-t})^2 + \cdots$$

$$= 1 + k\cdot 2^{-t}\left(1 + \frac{k-1}{2!}(2^{-t}) + \frac{(k-1)(k-2)}{3!}(2^{-t})^2 + \cdots\right)$$

$$< 1 + k\cdot 2^{-t}\left(1 + \frac{1}{2!}(0.1) + \frac{1}{3!}(0.1)^2 + \cdots\right)$$

$$= 1 + k\cdot 2^{-t}\cdot\frac{1}{0.1}(e^{0.1}-1) < 1 + (1.06)k\cdot 2^{-t}.$$

类似可得

$$(1-2^{-t})^k \geqslant 1 - (1.06)k\cdot 2^{-t}.$$

令 $$(1.06)2^{-t} = 2^{-t_1},$$

则 $$t_1 = t - \log_2(1.06) \doteq t - 0.084.$$

于是 $(1+2^{-t})^k<1+k\cdot 2^{-t_1}$, $(1-2^{-t})^k>1-k\cdot 2^{-t_1}$

其中 $t_1=t-0.084$

所以(2.2.10)的第2式可写成

$$1-(1.06)(n-k+1)2^{-t}<1+\eta_k<1+(1.06)(n-k+1)2^{-t}, \quad (2.2.12)$$

$$|\eta_k|<(1.06)(n-k+1)2^{-t}=(n-k+1)2^{-t_1}.$$

综上可得

$$\begin{cases} s_n=fl(x_1+x_2+\cdots+x_n)=x_1(1+\eta_1)+\cdots \\ \qquad +x_n(1+\eta_n), \\ |\eta_k|\leqslant(n-k+1)\cdot 2^{-t_1} \quad (k=1,2,\cdots,n). \end{cases} \quad (2.2.13)$$

(2.2.10)，(2.2.13)表明，连加的计算和是带有相对误差的各数$x_k(1+\eta_k)$的精确和，诸相对误差限的大小，因参加运算的各数的先后次序而异，其先参加运算者在和式中引起的误差的增长也较大。因此，求和时应先安排数值较小的数参加运算，这样做，在多数的情况下能取得较高的精度。

(2) **连乘**

设 $p_1=x_1$,

$$p_2=fl(p_1x_2)\equiv(p_1x_2)(1+\varepsilon_2).$$

$$\vdots$$

$$p_k=fl(p_{k-1}x_k)\equiv(p_{k-1}x_k)(1+\varepsilon_k),$$

$$|\varepsilon_k|\leqslant 2^{-t} \qquad (k=2,3,\cdots,n),$$

则有

$$\begin{cases} p_n=fl(x_1x_2\cdots x_n)=x_1x_2\cdots x_n(1+e), \\ (1-2^{-t})^{n-1}\leqslant 1+e\leqslant(1+2^{-t})^{n-1}, \end{cases} \quad (2.2.14)$$

由于 $p_n=x_1x_2\cdots x_n(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n)$，且 $|\varepsilon_k|\leqslant 2^{-t}$，若令

$$(1+\varepsilon_2)(1+\varepsilon_3)\cdots(1+\varepsilon_n)=1+e,$$

则有

$$(1-2^{-t})^{n-1} \leqslant 1+e \leqslant (1+2^{-t})^{n-1}. \qquad (2.2.15)$$

其上、下界与连加不一样，它与计算次序无关。

2.1.5 内积

设 x_i，$y_i (i=1,2,\cdots,n)$ 均为浮点规格化的数，求 $s_n = fl(x_1y_1 + x_2y_2 + \cdots + x_ny_n)$。计算次序是先求出两两乘积的单字长浮点数，再依次相加，直到求出最后结果。即设

$$p_1 = fl(x_1y_1) = s_1,$$

$$p_2 = fl(x_2y_2), \quad s_2 = fl(s_1 + p_2),$$

$$\vdots$$

$$p_k = fl(x_ky_k), \quad s_k = fl(s_{k-1} + p_k),$$

$$k = 2, 3, \cdots, n,$$

则有

$$\begin{cases} s_n = x_1y_1(1+\varepsilon_1) + x_2y_2(1+\varepsilon_2) + \cdots + x_ny_n(1+\varepsilon_n), \\ (1-2^{-t})^{n-k+2} \leqslant 1+\varepsilon_k \leqslant (1+2^{-t})^{n-k+2}, \qquad (2.2.16) \\ \quad k = 1, 2, \cdots, n, \end{cases}$$

因为 $\quad p_k \equiv x_ky_k(1+\xi_k), \quad |\xi_k| \leqslant 2^{-t},$

$$s_k \equiv (s_{k-1} + p_k)(1+\eta_k), \quad |\eta_n| \leqslant 2^{-t},$$

所以

$$s_n = x_1y_1(1+\varepsilon_1) + x_2y_2(1+\varepsilon_2) + \cdots + x_ny_n(1+\varepsilon_n).$$

其中

$$1+\varepsilon_1 = (1+\xi_1)(1+\eta_2)\cdots(1+\eta_n),$$

$$1+\varepsilon_2 = (1+\xi_2)(1+\eta_2)\cdots(1+\eta_n),$$

$$\vdots$$

$$1+\varepsilon_k = (1+\xi_k)(1+\eta_k)\cdots(1+\eta_n),$$

$$k = 3, 4, \cdots, n.$$

从而

$$\begin{cases} (1-2^{-t})^{n} \leqslant 1+\varepsilon_1 \leqslant (1+2^{-t})^{n}, \\ (1-2^{-t})^{n-k+2} \leqslant 1+\varepsilon_k \leqslant (1+2^{-t})^{n-k+2}, \\ \quad k=2,3,\cdots,n. \end{cases} \tag{2.2.17}$$

(2.2.17)还可统一地写成

$$(1-2^{-t})^{n-k+2} \leqslant 1+\varepsilon_k \leqslant (1+2^{-t})^{n-k+2} \tag{2.2.18}$$

$$k=1,2,\cdots,n.$$

当 $k \cdot 2^{-t} < 0.1$ 时，据(2.2.12)有

$$1-(1.06)(n-k+2)2^{-t} < 1+\varepsilon_k < 1+(1.06)(n-k+2)2^{-t},$$

即

$$|\varepsilon_k| < (n-k+2)(1.06)2^{-t} = (n-k+2)2^{-t_1}. \tag{2.2.19}$$

其中 $t_1 = t - 0.084$ 由此得到

$$\begin{cases} s_n = x_1y_1(1+\varepsilon_1) + x_2y_2(1+\varepsilon_2) + \cdots \\ \qquad + x_ny_n(1+\varepsilon_n), \\ |\varepsilon_k| \leqslant (n-k+2)2^{-t_1} \quad k=1,2,\cdots,n. \end{cases} \tag{2.2.20}$$

以上这些结果为算法的浮点误差分析及稳定性讨论提供了良好的基础。

2.2 消去法的浮点误差分析

利用前面的浮点误差分析的基本知识，可以对求解线性方程组的消元法进行误差分析，在此仅对列主元素法进行分析，全主元法可以类似地进行。

在做误差分析时，为叙述简单起见，设消元是按自然次序进行的，即每次消元时不需要进行行的交换，按自然次序选取的主元就是列主元。设求解的线性方程组为

$$A\boldsymbol{x} = \boldsymbol{b}.$$

首先考虑对 A 进行LR分解。设 $A = A^{(1)}$，各步消元后产生的计算输出矩阵依次为 $A^{(2)}$，$A^{(3)}$，$\cdots$，$A^{(n)} = \bar{R} = (r_{ij})_{n\times n}$，其中$A^{(k)}$的元素为$a_{ij}^{(k)}$，而且当 $j<k$，$i>j$时 $a_{ij}^{(k)}=0$，即$A^{(k)}$在

前$(k-1)$列的对角元以下的元素均为0。由普通消去法对$A^{(k)}$再进行一步消元后得$A^{(k+1)}$，其$a_{ij}^{(k+1)}$应为

$$a_{ij}^{(k+1)}=\begin{cases}0, & i\geqslant k+1,\ j=k,\\ fl(a_{ij}^{(k)}-m_{ik}\cdot a_{kj}^{(k)}), & i\geqslant k+1,\ j\geqslant k+1,\\ a_{ij}^{(k)}, & \text{其他},\end{cases} \quad (2.2.21)$$

其中
$$m_{ik}=fl(a_{ik}^{(k)}/a_{kk}^{(k)}) \quad i\geqslant k+1.$$

由于

$$m_{ik}=fl(a_{ik}^{(k)}/a_{kk}^{(k)})=\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}(1+\varepsilon_{ik}),\quad |\varepsilon_{ik}|\leqslant 2^{-t}. \quad (2.2.22)$$

于是当$i\geqslant k+1$，$j=k$时在(i,k)处的元素的精确结果应为

$$\bar{a}_{ik}^{(k+1)}=a_{ik}^{(k)}-m_{ik}a_{kk}^{(k)}=a_{ik}^{(k)}-\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}(1+\varepsilon_{ik})a_{kk}^{(k)}$$

$$=-a_{ik}^{(k)}\varepsilon_{ik}. \quad (2.2.23)$$

这个结果一般不为零，当按消去法要求，强使该处的输出为零时，需要引入一个平衡的误差$\varepsilon_{ik}^{(k)}$使$a_{ik}^{(k+1)}=\bar{a}_{ik}^{(k+1)}+\varepsilon_{ik}^{(k)}=0$，此时应有

$$\varepsilon_{ik}^{(k)}=a_{ik}^{(k)}\varepsilon_{ik}. \quad (2.2.24)$$

当$i\geqslant k+1$，$j\geqslant k+1$时

$$a_{ij}^{(k+1)}=fl(a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)})$$
$$=[a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}(1+\varepsilon_{ij})](1+\eta_{ij}), \quad (2.2.25)$$
$$|\varepsilon_{ij}|\leqslant 2^{-t},\quad |\eta_{ij}|\leqslant 2^{-t}.$$

从而有

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}+[a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}(1+\varepsilon_{ij})]\eta_{ij}$$
$$-m_{ik}a_{kj}^{(k)}\varepsilon_{ij}$$
$$=a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}+\frac{a_{ij}^{(k+1)}}{1+\eta_{ij}}\eta_{ij}-m_{ik}a_{kj}^{(k)}\varepsilon_{ij}. \quad (2.2.26)$$

改写(2.2.26)为

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}+\varepsilon_{ij}^{(k)} \quad (2.2.27)$$

其中

$$\varepsilon_{ij}^{(k)}=\frac{a_{ij}^{(k+1)}}{1+\eta_{ij}}\eta_{ij}-m_{ik}a_{kj}^{(k)}\varepsilon_{ij}\quad(i,j\geqslant k+1),$$

因为(2.2.21)的第三式不进行运算，所以没有引进误差，即

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}+\varepsilon_{ij}^{(k)},\ \ \varepsilon_{ij}^{(k)}=0。\tag{2.2.28}$$

综合(2.2.24)、(2.2.27)、(2.2.28)得到

$$a_{ij}^{(k+1)}=\begin{cases}a_{ij}^{(k)}-m_{ik}a_{kk}^{(k)}+\varepsilon_{ij}^{(k)}, & i\geqslant k+1,\ j=k,\\ a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}+\varepsilon_{ij}^{(k)}, & i\geqslant k+1,\ j\geqslant k+1,\\ a_{ij}^{(k)}+\varepsilon_{ij}^{(k)}, & \text{其他},\end{cases}\tag{2.2.29}$$

其中

$$\varepsilon_{ij}^{(k)}=\begin{cases}a_{ik}^{(k)}\varepsilon_{ik},\ \ |\varepsilon_{ik}|\leqslant 2^{-t},\ \ i\geqslant k+1,\ \ j=k,\\ \dfrac{a_{ij}^{(k+1)}}{1+\eta_{ij}}\eta_{ij}-m_{ik}a_{kj}^{(k)}\varepsilon_{ij},\ \ |\eta_{ij}|,|\varepsilon_{ij}|\leqslant 2^{-t},\\ i\geqslant k+1,\ \ j\geqslant k+1,\\ 0,\ \ \text{其他}。\end{cases}\tag{2.2.30}$$

设$E^{(k)}$是以$\varepsilon_{ij}^{(k)}$为元素的误差矩阵，则(2.2.29)可以写成

$$A^{(k+1)}=A^{(k)}-M_kA^{(k)}+E^{(k)},\tag{2.2.31}$$

其中

$$M_k=\begin{pmatrix}0 & & & & & \\ & \ddots & & & & \\ \cdot & & 0 & & & \\ \cdot & & m_{k+1,k} & 0 & & \\ \cdot & & m_{k+2,k} & & 0 & \\ & & \vdots & & & \ddots \\ 0 & & m_{n,k} & & & & 0\end{pmatrix}。\tag{2.2.32}$$

令 $k=1,2,\cdots,n-1$，相应地得到

$$A^{(2)}=A^{(1)}-M_1A^{(1)}+E^{(1)},$$

$$A^{(3)}=A^{(2)}-M_2A^{(2)}+E^{(2)},$$

$$\vdots$$

$$A^{(n)}=A^{(n-1)}-M_{n-1}A^{(n-1)}+E^{(n-1)}。$$

两边求和，得到

$$A^{(n)}=A^{(1)}-\sum_{i=1}^{n-1}M_iA^{(i)}+\sum_{i=1}^{n-1}E^{(i)}. \qquad (2.2.33)$$

注意 $A^{(k)}$ 的第 k 行与 $A^{(k+1)},\cdots,A^{(n)}$的第 k 行元素是一样的。而从（2.2.32）又可以看出，$M_kA^{(k)}$ 只与 $A^{(k)}$ 的第 k 行元素有关，因此

$$M_kA^{(k)}=M_kA^{(n)}=M_k\tilde{R},$$

于是(2.2.33)可写成

$$\tilde{R}=A^{(1)}-(\sum_{i=1}^{n-1}M_i)\tilde{R}+\sum_{i=1}^{n-1}E^{(i)},$$

经整理得

$$(M_1+M_2+\cdots+M_{n-1}+I)\tilde{R}=A+\sum_{i=1}^{n-1}E^{(i)}. \qquad (2.2.34)$$

令

$$M_1+M_2+\cdots+M_{n-1}+I=\tilde{L}=(l_{ij})_{n\times n},$$

$$E^{(1)}+E^{(2)}+\cdots+E^{(n-1)}=E,$$

则

$$\tilde{L}\tilde{R}=A+E, \qquad (2.2.35)$$

其中 $\tilde{L}$ 是单位下三角阵，$\tilde{R}$是上三角阵。从而我们得到了A的带有误差矩阵E的矩阵 $A+E$ 的精确的LR分解。

下面讨论E的元素的大小，为此，首先需要估计E的元素中的量m_{ik}和 $a_{ij}^{(k)}$ 的界。

由于假定按自然次序的消去法就是列主元消去法，所以$|a_{kk}^{(k)}|\geqslant|a_{ik}^{(k)}|$ $(i\geqslant k+1)$，且$m_{ik}=fl(a_{ik}^{(k)}/a_{kk}^{(k)})$，因此一般有

$$|m_{ik}|\leqslant 1,\ \text{即}|l_{ik}|\leqslant 1.$$

令

$$\rho=\max_{i,j,k}|a_{ij}^{(k)}|/\|A\|_\infty. \qquad (2.2.36)$$

则对任何$1\leqslant i,j,\ k\leqslant n$，均有

$$|a_{ij}^{(k)}|\leqslant\rho\|A\|_\infty \qquad (2.2.37)$$

及 $$|r_{ij}| \leqslant \rho\|A\|_\infty,$$

其中r_{ij}是上三角阵$\widetilde{\mathbf{R}}$的(i,j)元素．这里虽然不能对ρ的值做先验估计，但它是可以在消元过程中求出的。有了m_{ik}和$a_{ij}^{(k)}$的界以后，就可以估计$\varepsilon_{ij}^{(k)}$的界了。由(2.2.30)的第1式有

$$|\varepsilon_{ij}^{(k)}| = |a_{ik}^{(k)}\varepsilon_{ik}| \leqslant \rho\|A\|_\infty \cdot 2^{-t},$$
$$i \geqslant k+1,\quad j=k.$$

由(2.2.30)的第2式，得到

$$\begin{aligned}|\varepsilon_{ij}^{(k)}| &= \left|\frac{a_{ij}^{(k+1)}\eta_{ij}}{1+\eta_{ij}} - m_{ik}a_{kj}^{(k)}\varepsilon_{ij}\right| \\ &\leqslant \frac{|a_{ij}^{(k+1)}||\eta_{ij}|}{|1+\eta_{ij}|} + |m_{ik}||a_{kj}^{(k)}||\varepsilon_{ij}| \\ &\leqslant \frac{\rho\|A\|_\infty \cdot 2^{-t}}{1-2^{-t}} + \rho\|A\|_\infty \cdot 2^{-t} \\ &\leqslant \rho\|A\|_\infty \cdot 2^{-t}\left(\frac{1}{1-2^{-t}}+1\right).\end{aligned}$$

由于

$$\frac{1}{1-2^{-t}} \doteq 1+2^{-t},$$

所以

$$|\varepsilon_{ij}^{(k)}| \leqslant \rho\|A\|_\infty \cdot 2^{-t}\begin{cases}1, & i\geqslant k+1,\ j=k, \\ 2+2^{-t}, & i\geqslant k+1,\ j\geqslant k+1, \\ 0, & \text{其他}.\end{cases} \tag{2.2.38}$$

于是得

$$|E^{(k)}| \leqslant \rho\|A\|_\infty \cdot 2^{-t}\left[\begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \hline & 1 & 2+2^{-t} & \cdots & 2+2^{-t} \\ 0 & \vdots & \vdots & & \vdots \\ & 1 & 2+2^{-t} & \cdots & 2+2^{-t}\end{array}\right] \begin{array}{c} \\ (\text{第}k+1\text{行}) \\ \\ \\ \end{array}$$

(第k列)

又由 $E=\sum_{i=0}^{n-1}E^{(i)}$ 所以

$$
|E|\leqslant\rho\|A\|_\infty\cdot 2^{-t}\begin{pmatrix}
0 & 0 & 0 & \cdots & 0 & 0\\
1 & 2 & 2 & \cdots & 2 & 2\\
1 & 3 & 4 & \cdots & 4 & 4\\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\
1 & 3 & 5 & \cdots & 2n-4 & 2n-4\\
1 & 3 & 5 & \cdots & 2n-3 & 2n-2
\end{pmatrix}
$$

$$
+2^{-t}\begin{pmatrix}
0 & 0 & 0 & \cdots & 0 & 0\\
0 & 1 & 1 & \cdots & 1 & 1\\
0 & 1 & 2 & \cdots & 3 & 3\\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\
0 & 1 & 2 & \cdots & n-2 & n-2\\
0 & 1 & 2 & \cdots & n-2 & n-1
\end{pmatrix},
$$

$$
\begin{aligned}
\|E\|_\infty=\||E|\|_\infty &\leqslant\rho\|A\|_\infty\cdot 2^{-t}\{[\sum_{i=1}^{n}(2i-1-1]+2^{-t}\sum_{i=1}^{n-1}i\}\\
&\leqslant\rho\|A\|_\infty\cdot 2^{-t}(n^2+2^{-t}\cdot n^2/2)\\
&\leqslant\rho\|A\|_\infty\cdot n^2\cdot 2^{-t+1}.
\end{aligned}
$$

定理2.2.1 使用列主元消去法时，实际计算出来的三角因子 $\widetilde{\mathrm{L}}$ 和 $\widetilde{\mathrm{R}}$ 满足

$$\widetilde{\mathrm{L}}\,\widetilde{\mathrm{R}}=A+E, \tag{2.2.39}$$

$$\|E\|_\infty\leqslant n^2\rho\|A\|_\infty 2^{-t+1}. \tag{2.2.40}$$

这个重要结果表明，对 A 的实际计算所得到的三角分解，可以看作是 $A+E$ 的精确的三角分解。对于一般矩阵，如果 ρ 不是很大，那末，在分解过程中由于舍入误差而引起的误差矩阵，其元素的界与机器精度几乎是同阶的。

完成三角分解的误差分析以后，对于求解方程组 $Ax=b$ 的

问题将归结为求解两个三角方程组的问题，即

$$\tilde{L}y=b \quad 和 \quad \tilde{R}x=y, \tag{2.2.41}$$

因此，要完成对消元法的误差分析，首先要估计求解三角方程组

$$Ux=b \tag{2.2.42}$$

中的解的误差。不失一般性，设U为下三角阵，于是(2.2.42)可以写成

$$\begin{cases} u_{11}x_1 = b_1, \\ u_{21}x_1+u_{22}x_2 = b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ u_{n1}x_1+u_{n2}x_2+\cdots+u_{nn}x_n=b_n, \end{cases} \tag{2.2.43}$$

依次解出x_1，x_2，…，x_n，且由(2.2.8)可得：

$$x_1=fl(b_1/u_{11})=b_1/(u_{11}(1+\varepsilon_n)),\quad |\varepsilon_n|<2^{-t}, \tag{2.2.44}$$

$$\begin{aligned} x_i &= fl\left(\frac{-u_{i1}x_1-u_{i2}x_2-\cdots-u_{ii-1}x_{i-1}+b_i}{u_{ii}}\right) \\ &= \frac{fl(-u_{i1}x_1-u_{i2}x_2-\cdots-u_{ii-1}x_{i-1}+b_i)}{u_{ii}(1+\varepsilon_{ii})} \\ &= \frac{fl(-u_{i1}x_1-u_{i2}x_2-\cdots-u_{ii-1}x_{i-1})+b_i}{u_{ii}(1+\varepsilon_{ii})(1+\varepsilon'_{ii})}. \end{aligned}$$

由(2.2.20)，(2.2.19)得到

$$x_i=\frac{-u_{i1}(1+\varepsilon_{i1})x_1-u_{i2}(1+\varepsilon_{i2})x_2-\cdots-u_{ii-1}(1+\varepsilon_{ii-1})x_{i-1}+b_i}{u_{ii}(1+\varepsilon_{ii})(1+\varepsilon'_{ii})}, \tag{2.2.45}$$

其中

$$\begin{cases} |\varepsilon_{ii}|,|\varepsilon'_{ii}|\leqslant 2^{-t},\quad i=1,2,\cdots,n, \\ |\varepsilon_{i1}|<(i-1)(1.06)2^{-t},\quad i=2,3,\cdots,n, \\ |\varepsilon_{ik}|<((i-1)-k+2)(1.06)2^{-t} \\ \qquad\qquad =(i+1-k)(1.06)2^{-t} \\ (i=2,3,\cdots,n,\quad k=2,3,\cdots,i-1). \end{cases} \tag{2.2.46}$$

由此可知

$$
\begin{aligned}
&u_{11}(1+\varepsilon_{11})x_1=b_1,\\
&u_{i1}(1+\varepsilon_{i1})x_1+u_{i2}(1+\varepsilon_{i2})x_2+\cdots\\
&\quad+u_{ii-1}(1+\varepsilon_{ii-1})x_{i-1}+u_{ii}(1+\varepsilon_{ii})(1+\varepsilon'_{ii})x_i=b_i,\\
&i=2,3,\cdots,n.
\end{aligned}
\tag{2.2.47}
$$

写成矩阵形式 $(U+\delta U)x=b$,

其中

$$
\delta U=\begin{pmatrix}
u_{11}\varepsilon_{11} & & & \\
u_{21}\varepsilon_{21} & u_{22}\delta_{22} & & \\
\vdots & & & \\
u_{n1}\varepsilon_{n1} & u_{n2}\varepsilon_{n2} & \cdots u_{nn-1}\varepsilon_{nn-1} & u_{nn}\delta_{nn}
\end{pmatrix}
$$

$$
\delta_{ii}=\varepsilon_{ii}+\varepsilon'_{ii}+\varepsilon_{ii}\varepsilon'_{ii},\quad i=1,2,\cdots,n,
$$

$$
|\delta_{ii}|\leqslant|\varepsilon_{ii}|+|\varepsilon'_{ii}|+|\varepsilon_{ii}\varepsilon'_{ii}|\leqslant 2.2^{-t}+2^{-2t}. \tag{2.2.48}
$$

由(2.2.46)，(2.2.48)可知

$$
|\delta U|\leqslant(1.06)2^{-t}
\begin{pmatrix}
|u_{11}| & & & \\
1|u_{21}| & 2|u_{22}| & & \\
2|u_{31}| & 2|u_{32}| & 2|u_{33}| & \\
\cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\
(n-1)|u_{n1}| & (n-1)|u_{n2}| & \cdots 2|u_{nn-1}| & 2|u_{nn}|
\end{pmatrix}
\tag{2.2.49}
$$

估计 δ_U 的范数 $\|\cdot\|_\infty$ 得到

$$
\begin{aligned}
\|\delta_U\|&\leqslant(1.06)2^{-t}\left(\sum_{i=1}^{n} i\right)\max_{i,k}|u_{ik}|\\
&=\frac{n^2+n}{2}(1.06)\cdot 2^{-t}\max_{i,k}|u_{ik}|.
\end{aligned}
\tag{2.2.50}
$$

定理2.2.2 系数矩阵为三角阵的线性方程组 $Ux=b$ 的计算解是方程组 $(U+\delta_U)x=b$ 的精确解。其中 δ_U 满足(2.2.49)和(2.2.50)。

将这个定理应用到求解三角方程组 $\tilde{L}\boldsymbol{y}=\boldsymbol{b}$ 和 $\tilde{R}\boldsymbol{x}=\boldsymbol{y}$，最后得到的计算解应满足

$$(\tilde{L}+\delta\tilde{L})(\tilde{R}+\delta\tilde{R})\boldsymbol{x}=\boldsymbol{b}$$

或
$$(\tilde{L}\tilde{R}+\delta\tilde{L}\cdot\tilde{R}+\tilde{L}\cdot\delta\tilde{R}+\delta\tilde{L}\cdot\delta\tilde{R})\boldsymbol{x}=\boldsymbol{b}.$$

将 $\tilde{L}\tilde{R}=A+E$ 代入上式，得

$$(A+E+\delta\tilde{L}\cdot\tilde{R}+\tilde{L}\cdot\delta\tilde{R}+\delta\tilde{L}\cdot\delta\tilde{R})\boldsymbol{x}=\boldsymbol{b}, \tag{2.2.51}$$

由(2.2.36)，(2.2.37)和(2.2.50)知

$$\begin{cases}\|\tilde{L}\|_\infty\leqslant n,\\ \|\tilde{R}\|_\infty\leqslant n\rho\|A\|_\infty,\\ \|\delta\tilde{L}\|_\infty\leqslant\dfrac{n^2+n}{2}(1.06)\cdot 2^{-t}\max\limits_{i,k}|l_{i,k}|\leqslant\dfrac{n^2+n}{2}(1.06)2^{-t},\\ \|\delta\tilde{R}\|_\infty\leqslant\dfrac{n^2+n}{2}(1.06)\cdot 2^{-t}\max\limits_{i,k}|r_{i,k}|\\ \qquad\leqslant\dfrac{n^2+n}{2}(1.06)\rho\|A\|_\infty 2^{-t}.\end{cases} \tag{2.2.52}$$

现令

$$\delta A=E+\delta\tilde{L}\cdot\tilde{R}+\tilde{L}\cdot\delta\tilde{R}+\delta\tilde{L}\cdot\delta\tilde{R}, \tag{2.2.53}$$

则
$$\|\delta A\|_\infty\leqslant\|E\|_\infty+\|\delta\tilde{L}\|_\infty\|\tilde{R}\|_\infty+\|\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty+\|\delta\tilde{L}\|_\infty\|\delta\tilde{R}\|_\infty,$$

由(2.2.40)，(2.2.52)可得

$$\|\delta A\|_\infty\leqslant 1.06n^2\rho\|A\|_\infty\cdot 2^{-t}\left(3+n+\frac{(n+1)^2}{2}2^{-t}\right).$$

又因为，在实际计算中，常有 $(n+1)^2\cdot 2^{-t}/2\ll 1$，故有

$$\|\delta A\|_\infty\leqslant 1.06(n^3+4n^2)\rho\cdot 2^{-t}\|A\|_\infty \tag{2.2.54}$$

定理2.2.3 用列主元消去法求出的线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 的计算解 $\boldsymbol{x}$ 是方程组

$$(A+\delta A)\boldsymbol{x}=\boldsymbol{b}$$

的精确解。其中误差矩阵 δA 由(2.2.53)给出，且满足估计式(2.2.54)。

从定理2.2.3的证明过程中，我们可以看到，由(2.2.54)所表示的误差矩阵的范数的上界估计式还是较保守的。如果令$\bar{A}=A+\delta A$，则定理2.2.3的结果表明$\phi^*(A,b)=\phi(\bar{A},b)$，其中$\phi^*(A,b)$表示$Ax=b$的计算解，$\phi(\bar{A},b)$表示$\bar{A}x=b$的精确解，而且有$\|\bar{A}-A\|_\infty\leqslant1.06(n^3+4n^2)\rho\cdot2^{-t}\|A\|_\infty$。只要$1.06(n^3+4n^2)\rho\|A\|_\infty$有一个不大的常数上界，即$|\delta A|$的元素很小，那末，计算解$\phi^*$(按列的部份主元素法)就是稳定的。在$1.06(n^3+4n^2)\rho\|A\|_\infty$中，需要进一步分析的是$\rho$的上界的估计问题。

由于

$$|a_{ij}^{(k)}|\leqslant|a_{ij}^{(k-1)}-m_{ik}a_{kj}^{(k-1)}|\leqslant|a_{ij}^{(k-1)}|+|a_{kj}^{(k-1)}|$$
$$\leqslant2\max_{i,j}|a_{ij}^{(k-1)}|,$$

从而有

$$\max_{i,j}|a_{ij}^{(k)}|\leqslant2\max_{i,j}|a_{ij}^{(k-1)}|\leqslant2^2\max_{i,j}|a_{ij}^{(k-2)}|\leqslant\cdots$$
$$\leqslant2^{k-1}\max_{i,j}|a_{ij}|\leqslant2^{k-1}\|A\|_\infty,$$

因此

$$\rho\leqslant m\cdot2^{n-1},\quad m=\max_{i,j}|a_{ij}|/\|A\|_\infty\quad(\text{常数})。$$

从而ρ有一个与矩阵阶数n有关的上界，其中2^{n-1}是一个变化相当快的增长因子。但是在实践中很少达到这个界，因此在实践中解线性方程组的列主元素法被认为是稳定的。

2.3 解的精度

定理2.2.3说明了非奇异线性方程组$Ax=b$的计算解x正好是方程组

$$(A+\delta A)x=b$$

的精确解，并且给出了误差矩阵δA的界为

$$\|\delta A\|_\infty<1.05(n^3+4n^2)\rho\cdot2^{-t}\|A\|_\infty.$$

如果令 $f(n)=1.06(n^3+4n^2)\rho$，这是一个依赖于具体计算细节的量，一般说来，其值不会很大，将$f(n)$代人上式得到

$$\|\delta A\|_\infty < f(n)\|A\|_\infty \cdot 2^{-t}. \tag{2.2.55}$$

由定理 2.1.2 的推论知道，如果设 $\bar{\boldsymbol{x}}$ 是 $A\boldsymbol{x}=\boldsymbol{b}$ 的精确解，且 $\|A^{-1}\|_\infty\|\delta A\|_\infty<1$，则有

$$\frac{\|\bar{\boldsymbol{x}}-\boldsymbol{x}\|_\infty}{\|\bar{\boldsymbol{x}}\|_\infty} \leqslant \frac{\|A^{-1}\|_\infty\|\delta A\|_\infty}{1-\|A^{-1}\|_\infty\|\delta A\|_\infty}.$$

因此，当

$$\begin{aligned}\|A^{-1}\|_\infty\|\delta A\|_\infty &< f(n)\|A^{-1}\|_\infty\|A\|_\infty \cdot 2^{-t} \\ &= f(n)\operatorname{cond}(A)2^{-t}<1\end{aligned} \tag{2.2.56}$$

时，就有

$$\frac{\|\bar{\boldsymbol{x}}-\boldsymbol{x}\|_\infty}{\|\bar{\boldsymbol{x}}\|_\infty} \leqslant \frac{f(n)\operatorname{cond}(A)2^{-t}}{1-f(n)\operatorname{cond}(A)2^{-t}}. \tag{2.2.57}$$

如果 $f(n)$ 和 $\operatorname{cond}(A)$ 均不太大，并且令 $f(n)\operatorname{cond}(A)\leqslant 2^p$，则

$$\frac{\|\bar{\boldsymbol{x}}-\boldsymbol{x}\|_\infty}{\|\bar{\boldsymbol{x}}\|_\infty} \leqslant \frac{2^{-t+p}}{1-2^{-t+p}}, \tag{2.2.58}$$

从而我们可以期望用 t 位精度求得的计算解，能达到大约有 $k=t-p$ 位有效数字。而且对 $\|\boldsymbol{x}\|_\infty$ 有如下的估计式

$$\left(1-\frac{2^{-k}}{1-2^{-k}}\right)\|\bar{\boldsymbol{x}}\|_\infty \leqslant \|\boldsymbol{x}\|_\infty \leqslant \frac{\|\bar{\boldsymbol{x}}\|_\infty}{1-2^{-k}}. \tag{2.2.59}$$

但是必须注意，如果条件(2.2.56)不满足时，我们就不能肯定 $\boldsymbol{x}$ 有任何精度，而且 $(A+\delta A)$ 甚至可能是奇异的。由于对 $\|\delta A\|_\infty$ 的上界的估计是非常保守的，因此，利用(2.2.57)来估计解的精确度时难免会产生过估现象，有时候这种过估还是很严重的。

由于$f(n)$和$\mathrm{cond}(A)$均有可能进行估计，从而在算得x以前就可以定出x的相对误差界，即先验误差界。(2.2.57)中的误差界即属于先验误差界。现在再来分析依赖于近似解x本身的误差界，即通常所称的事后误差界。观察事后误差界的一个很自然的方法是把它代入方程组$Ax=b$，然后再看残向量

$$r=b-Ax \qquad (2.2.60)$$

范数的大小。对于x，如果$r=0$，则x就是$Ax=b$的精确解。然而，是否$\|r\|$较小时，近似解x的精确度就一定高呢？我们可以从下面的定理得出应有的结论。

定理2.2.4 设A是非奇异阵且$Ax=b\neq 0$，x是已知的近似解，$\bar{x}$为精确解，则

$$\frac{\|x-\bar{x}\|}{\|\bar{x}\|}\leqslant \mathrm{cond}(A)\frac{\|r\|}{\|b\|}. \qquad (2.2.61)$$

证明 因为$r=b-Ax$，且A非奇异，所以

$$A^{-1}r=A^{-1}b-x=\bar{x}-x,$$

于是

$$\|\bar{x}-x\|=\|A^{-1}r\|\leqslant\|A^{-1}\|\|r\|. \qquad (2.2.62)$$

又

$$\|b\|=\|A\bar{x}\|\leqslant\|A\|\|\bar{x}\|$$

故

$$\|\bar{x}\|\geqslant\|b\|/\|A\|. \qquad (2.2.63)$$

(2.2.62)除以(2.2.63)得

$$\frac{\|\bar{x}-x\|}{\|\bar{x}\|}\leqslant\frac{\|A^{-1}\|\|A\|\|r\|}{\|b\|}=\mathrm{cond}(A)\frac{\|r\|}{\|b\|}.$$

这个结果表明，解的精度不仅依赖于残向量的模$\|r\|$的大小，而且还依赖于矩阵的条件数。如果A是病态的，即使残向量的

模 $\|r\|$ 很小，也不能保证近似解的精度很高，而且对于精度高的近似解也可能有大的 $\|r\|$。

例2.2.2 方程组

$$\begin{bmatrix}1.000 & 1.001\\1.000 & 1.000\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}2.001\\2.000\end{bmatrix}$$

的精确解显然是 $\bar{x}=(1,1)^T$。向量 $x=(2,0)^T$ 并不接近 $\bar{x}$，而残向量

$$r=\begin{bmatrix}2.001\\2.000\end{bmatrix}-\begin{bmatrix}1.000 & 1.001\\1.000 & 1.000\end{bmatrix}\begin{bmatrix}2\\0\end{bmatrix}=\begin{bmatrix}0.001\\0.000\end{bmatrix}$$

却很小。此例说明残向量很小，但近似解的精度并不一定高。

方程组

$$\begin{bmatrix}1.000 & 1.001\\1.000 & 1.000\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix}$$

的精确解显然是 $\bar{x}=[-1000,1000]^T$，但是向量 $(-1001,1000)^T$ 虽然较接近于 $\bar{x}$，相应的残向量

$$\begin{bmatrix}1\\0\end{bmatrix}-\begin{bmatrix}1.000 & 1.001\\1.000 & 1.000\end{bmatrix}\begin{bmatrix}-1001\\1000\end{bmatrix}=\begin{bmatrix}1\\1\end{bmatrix}=r$$

却比 b 还大。此例说明近似解的精度虽然较高，但是残向量并不一定很小。

什么原因引起方程组病态以及怎样度量一个方程组是否病态呢？这一直是人们关心的一个困难的课题。Turing在1948年第一个使用了“条件数”这个术语。接着，不少数学工作者相继提出用条件数来度量相应矩阵的病态的程度，但是对条件数的大小的估计还缺少有效的方法。Noble 1969 年所给出的检查病态条件的各种方法的图表，使用起来仍不够理想。随着电子计算机的发展，解病态问题也变得愈来愈突出，近年来，除解

病态方程组的直接法和迭代法仍在继续发展外，还出现了预处理法，这是一种较有发展前途的方法，但是目前还都有一些局限性。因此，判别病态问题和解病态问题都是目前计算数学的重要课题。

关于算法的误差分析，经 Wilkinson 的努力，自1957年取得了比较大的进展后，直到目前向后误差分析还是分析舍入误差的最完整最系统的方法。近年来，它的概念更加系统，分析过程趋于简化，结果也更为精确而实用。

第二章 习 题

2.1 证明方程组

$$\begin{cases} x+y=2, \\ \dfrac{1}{6}x+\dfrac{1}{6}y=\dfrac{1}{3}, \\ x+2y=3 \end{cases}$$

有唯一解，但是它不适定。

2.2 设 $A\in R^{n\times n}$ 为 n 阶矩阵，$x=(x_1,x_2,\cdots,x_n)^T$，其中 a_{ij} 和 x_i 均为规格化浮点数，k 为某实数，求 $fl(kA)$，$fl(Ax)$。

2.3 $A,B\in R^{n\times n}$ 其中 a_{ij} 和 b_{ij} 均为规格化浮点数，求 $fl(A\cdot B)$，$fl(A+B)$。

2.4 设 x，y，z 都是具有 t 个数字的规格化浮点数，用例子说明

$$fl(fl(x+y)+z)\neq fl(x+fl(y+z)),$$

$$fl(fl(xy)z)=fl(xfl(yz))$$

是否一定成立？为什么？

2.5 设 $fl(x_1+\cdots+x_k)\equiv fl(fl(x_1+\cdots+x_{k-1})+x_k)$

$$(k=2,\cdots,n),$$

证明 $fl(\alpha_1+\alpha_2+\cdots+\alpha_n)=\alpha_1(1+\rho_1)^{n-1}+$

$$\alpha_2(1+\rho_2)^{n-1}+\alpha_3(1+\rho_3)^{n-2}+\cdots+\alpha_n(1+\rho_n),$$

其中 $|\rho_i|\leqslant u10^{-t}$，$u$是与1同阶的数。

2.6 设$x=39.25$，$y=0.23$，试将x，y分别化成$t=6$的规格化浮点数的机器表示形式$a=2^{b}a_1$和$b=2^{b_1}a_2$，并求$fl(a+b)$，$fl\left(\frac{a}{b}\right)$.

2.7 设二次方程

$$x^2-6.433x+0.009474=0,$$

试用求根公式计算该方程的最小根的计算值。

2.8 举出"病态"和不稳定算法的数值例子各一个。

2.9 如果系数矩阵A是三对角阵，使用部分主元素法求解$Ax=b$时，估计$\rho=\max\limits_{i,j,k}|a_{ij}^{(k)}|/\|A\|_\infty$的上界。

2.10 设A为实对称矩阵，R为正交阵，求$fl(R^TAR)$，并给出误差矩阵在某种范数下的上界。

2.11 证明定理2.1.1即证明关系式$k^{(1)}(A)\leqslant\text{cond}(A)\leqslant\sqrt{n}\,2^{n-1}k^{(1)}(A)$。

2.12 作镜像变换的误差分析。

参 考 书

[1] J. H. Wilkinson, "*The Algerbaic eigenvalue problems*", Oxford Univ. press, London and New York, 1965.

[2] Stewart. G, "*Introduction to Motrix Computation*", Academic press, New York 1973（中译本，矩阵计算引论，王国荣等译，上海科技出版社）。

[3] 曹志浩等编，矩阵计算和方程求根，1979．人民教育出版社．

[4] G. M. phillips, P. J. Taylor, "*Theory and Applications of Numerical Analysis*", Academic press, 1973（中译本，数值分析的理论及其应用，熊西文等译，上海科技出版社）．

第三章 特征值计算

§1 引 言

1.1 相似变换

用矩阵变换方法求矩阵的特征值，其基本思想是根据相似矩阵必有相同的特征谱这一性质，把矩阵化为形式简单的矩阵，使新矩阵的特征值便于计算。上三角阵是形式简单的矩阵，且其对角元就是它的特征值。

如果只限于用实数进行运算，那末，有复特征值的实矩阵在任何相似变换下都不能化为上三角阵。比上三角阵范围更广且特征值又易于计算的一类形式简单的矩阵是形如(3.1.1)的被称为拟上三角阵的矩阵：

$$\begin{pmatrix} \times & \times & & & & & & * \\ \times & \times & & & & & & \\ & & \ddots & & & & & \\ & & & \times & \times & & & \\ & & & \times & \times & & & \\ & & & & & \times & & \\ & & & & & & \ddots & \\ & & & & & & & \times \\ & & & & & & & & \times \end{pmatrix} \tag{3.1.1}$$

这种矩阵的下三角部分，除次对角线上有非零元素外，其余全为零，而次对角线上的两个非零元素之间必有一些零元，这是一个对角线上由一阶或二阶块组成的阵。而矩阵特征值就是这些块阵的特征值组成的。于是，求给定矩阵的特征值问题，就

可归结为：把给定矩阵用一系列相似变换化为上三角阵或拟上三角阵。

这个过程，通常是不能在有限步内结束的。例如，形如

$$F=\begin{pmatrix} -p_1 & -p_2 & \cdots & -p_n \\ 1 & 0 & & \\ & 1 & \ddots & \\ & & \ddots & \\ & & 1 & 0 \end{pmatrix} \tag{3.1.2}$$

的所谓 Frobenious 阵，其特征多项式为

$$f(\lambda)=\lambda^n+p_1\lambda^{n-1}+\cdots+p_{n-1}\lambda+p_n,$$

当 $n>5$ 时，它的零点，即 F 的特征值、一般是不能用 $p_i(i=1, 2, \cdots, n)$ 经有限次算术运算求出的。由此可知，使 $P^{-1}FP$ 化为 (3.1.1)的 P 也不能由 $p_i(i=1,2,\cdots,n)$ 经有限次运算求得。

使任意实矩阵 A 化为拟三角阵的相似变换矩阵是否存在与这个变换矩阵是否可经由有限步运算得到，这是两个不同的问题，后者已予说明，前者可用下列定理来回答。

定理 3.1.1　设 A 是实矩阵，则存在非奇异的实矩阵 P，使 $P^{-1}AP$ 成为拟上三角阵。

证明　用归纳法，当 $n=1$ 时，定理显然成立。设当 $n\leqslant k$ 时定理也成立。则当 $n=k+1$ 时，任取 A 的一个特征值 λ，当 λ 是实数时，可设其相应于 λ 的特征向量为 x，且 x 也是实的，这时，可以找到一个实的初等 Hermite 矩阵 H_1，此时 H_1 是实的正交阵，使

$$H_1x=\rho e_1,\quad |\rho|=\|x\|\neq 0 \tag{3.1.3}$$

由于

$$Ax=\lambda x$$

两端用 H_1 左乘，并注意 $\boldsymbol{x}=PH_1^{-1}\boldsymbol{e}_1$，得

$$H_1AH_1^{-1}\boldsymbol{e}_1=\lambda\boldsymbol{e}_1, \tag{3.1.4}$$

可见 $H_1AH_1^{-1}$ 具有下列形状.

$$H_1AH_1^{-1}=\left(\begin{array}{c|c}\lambda & \boldsymbol{b}_1^T \\ \hline \boldsymbol{o} & A_{n-1}\end{array}\right). \tag{3.1.5}$$

据归纳法假设，A_{n-1} 是阶数 $\leqslant k$ 的阵，存在正交阵 H_2 使 $H_2A_{n-1}H_2^{-1}$ 具有(3.1.1)的形状，取

$$P=\left(\begin{array}{c|c}1 & \boldsymbol{o} \\ \hline \boldsymbol{o} & H_2\end{array}\right)H_1, \tag{3.1.6}$$

则 PAP^{-1} 亦具有(3.1.1)的形状。且 P 是非奇异阵。

当 λ 是复数时，令 $\lambda=\alpha+i\beta$，$\boldsymbol{x}=\boldsymbol{u}+i\boldsymbol{v}$，而$\alpha$、$\beta$、$\boldsymbol{u}$、$\boldsymbol{v}$、都是实的，且 $\beta\neq 0$，由 $A\boldsymbol{x}=\lambda\boldsymbol{x}$，得

$$\begin{cases}A\boldsymbol{u}=\alpha\boldsymbol{u}-\beta\boldsymbol{v},\\ A\boldsymbol{v}=\beta\boldsymbol{u}+\alpha\boldsymbol{v},\end{cases} \tag{3.1.7}$$

可以证明 $\boldsymbol{u}$、$\boldsymbol{v}$ 是线性无关的。不然，设 p、q 不全为零，而 $p\boldsymbol{u}+q\boldsymbol{v}=0$，则由(3.1.7)，可得

$$\beta(-p\boldsymbol{v}+q\boldsymbol{u})=0,$$

从而有

$$p\boldsymbol{u}+q\boldsymbol{v}=0 \quad 及 \quad q\boldsymbol{u}-p\boldsymbol{v}=0.$$

由于 $p^2+q^2\neq 0$，可得 $\boldsymbol{u}=\boldsymbol{v}=0$，矛盾。故 $\boldsymbol{u}$、$\boldsymbol{v}$ 线性无关. 这时，可以找到初等正交阵 H_1，使

$$H_1(\boldsymbol{u},\boldsymbol{v})=\begin{pmatrix}r_{11} & r_{12}\\ 0 & r_{22}\\ 0 & 0\\ \vdots & \vdots\\ 0 & 0\end{pmatrix}, \quad r_{11}r_{22}\neq 0, \tag{3.1.8}$$

据(3.1.7)，有

$$A(u,v)=(u,v)\begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix} \tag{3.1.9}$$

及

$$H_1AH_1^{-1}\begin{pmatrix}r_{11} & r_{12}\\ 0 & r_{22}\\ \vdots & \vdots\\ 0 & 0\end{pmatrix}=\begin{pmatrix}r_{11} & r_{22}\\ 0 & r_{22}\\ \vdots & \vdots\\ 0 & 0\end{pmatrix}\begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix}. \tag{3.1.10}$$

比较上式两端，知 $H_1AH_1^{-1}$ 具有下列形状

$$H_1AH_1^{-1}=\begin{pmatrix}A_2 & B_2\\ 0 & A_{n-2}\end{pmatrix}. \tag{3.1.11}$$

这里

$$A_2=\begin{pmatrix}r_{11} & r_{12}\\ 0 & r_{22}\end{pmatrix}\begin{pmatrix}\alpha & -\beta\\ \beta & \alpha\end{pmatrix}\begin{pmatrix}r_{11} & r_{12}\\ 0 & r_{22}\end{pmatrix}^{-1},$$

可见 A_2 的特征值仍为 $\alpha\pm i\beta$，由于 A_{n-2} 的阶是 $k-1$，据归纳法假设，仿前面的讨论，知定理对 $n=k+1$ 也成立。

定理3.1.1的证明，将以A的特征值及特征向量为前提，并不能看作是使 $P^{-1}AP$ 成为拟三角阵的P的存在的构造性证明，但定理本身却为P的构造性算法提供了理论根据。

1.2 LR 算法

设实矩阵A有三角分解

$$A=LR, \tag{3.1.12}$$

其中L是单位下三角阵，R是上三角阵，L显然有逆，且 $L^{-1}AL$ 是A的一个相似矩阵，记

$$A_2=L^{-1}A_1L,\quad A_1=A, \tag{3.1.13}$$

代入$A=LR$，则

$$A_2=RL. \tag{3.1.14}$$

它不过是A的三角分解中两个矩阵按相反的次序的乘积，A_2是新矩阵，用它取代$A_1=A$的地位，并重复刚才的过程，又可得到A_3，仿此，便可得到与A相似的矩阵序列：

$$A_1=A,\quad A_s=L_sR_s,\quad A_{s+1}=R_sL_s,$$
$$s=1,2,\cdots \tag{3.1.15}$$

这里每个L_s都是单位下三角阵，每个R_s都是上三角阵。

注意，并非每个矩阵A，甚至非奇异的矩阵A都有LR分解，例如

$$A=\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$$

是非奇异的，但却没有LR分解。

当A非奇异时，据第一章，A有LR分解的充要条件应是A的各阶首主子式不为零，在此条件下分解将是唯一的。本节在讨论A的LR分解时，我们总是假定A及其LR分解序列中的每一矩阵A_s的各阶首主子式均不为零。

设$\{A_s\}$，$s=1,2,\cdots$是A的LR分解序列，则可证：

$$A_1=L_1L_2\cdots L_{s-1}A_sL_{s-1}^{-1}\cdots L_2^{-1}L_1^{-1}, \tag{3.1.16}$$
$$A_1L_1\cdots L_{s-1}=L_1\cdots L_{s-1}A_s. \tag{3.1.17}$$

记

$$T_s=L_1L_2\cdots L_s,\quad U_s=R_s\cdots R_1 \tag{3.1.18}$$

则有
$$A_1^s=T_sU_s, \tag{3.1.19}$$

这表明T_sU_s是A_1^s的LR分解。

1.3 LR 算法的收敛性

由$A_1^s=T_sU_s$及A_1的非奇异性，知A_1^s的三角分解应是唯一的。记$t_{ij}^{(s)}$为T_s中的(i,j)元素，其中$t_{ii}^{(s)}=1$，$t_{ij}^{(s)}=0$，$(j>i)$，令$\Delta_j^{(s)}$表示A_1^s的j阶首主子式，$\Delta_{ij}^{(s)}$表示$\Delta_j^{(s)}$中把第j行的元素换成A_1^s的第i行$(i>j)$上相应的元素后所成的

j 阶行列式，则有

$$t_{ij}^{(s)} = \Delta_{ij}^{(s)}/\Delta_j^{(s)} \qquad (i>j)。 \qquad (3.1.20)$$

下面考查 A 在什么条件下，其 LR 序列是收敛的。

设 A 有依模不相等的特征值。

$$|\lambda_1|>|\lambda_2|>\cdots>|\lambda_n|, \qquad (3.1.21)$$

对应的特征向量为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n$，记

$$\mathbf{X}=(\boldsymbol{x}_1,\cdots\boldsymbol{x}_n)=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix}, \qquad (3.1.22)$$

记 $\mathbf{X}$ 的逆阵为 Y：

$$\mathbf{X}^{-1}=\mathbf{Y}(\boldsymbol{y}_1, \cdots\boldsymbol{y}_n)=\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \vdots & y_{nn} \end{pmatrix}, \qquad (3.1.23)$$

于是有

$$A=X\mathrm{diag}(\lambda_1,\cdots,\lambda_n)Y,$$

$$A^s=X\mathrm{diag}(\lambda_1^s,\cdots,\lambda_n^s)Y,$$

及

$$(A^s)_{i,j}=\sum_{k=1}^{n} x_{ik}y_{kj}\lambda_k^s, \qquad (3.1.24)$$

A^s 的 p 阶首主子式为

$$\Delta_p^{(s)}=\begin{vmatrix} \sum_{k=1}^{n} x_{1k}y_{k1}\lambda_k^s, & \sum_{k=1}^{n} x_{1k}y_{k2}\lambda_k^s, & \cdots & \sum_{k=1}^{n} x_{1k}y_{kp}\lambda_k^s \\ \sum_{k=1}^{n} x_{2k}y_{k1}\lambda_k^s, & \sum_{k=1}^{n} x_{2k}y_{k2}\lambda_k^s, & \cdots & \sum_{k=1}^{n} x_{2k}y_{kp}\lambda_k^s \\ \vdots & \vdots & & \vdots \\ \sum_{k=1}^{n} x_{pk}y_{k1}\lambda_k^s, & \sum_{k=1}^{n} x_{pk}y_{k2}\lambda_k^s, & \cdots & \sum_{k=1}^{n} x_{pk}y_{kp}\lambda_k^s \end{vmatrix} \qquad (3.1.25)$$

$$=X_p\bar{Y}_p(\lambda_1,\cdots,\lambda_p)^S+\cdots$$

这里 X_p、Y_p 分别表示X、Y 的 p 阶首主子式，上式中没有明显写出的项是形如

$$X_{k_1\cdots k_p}\overline{Y}_{k_1\cdots k_p}(\lambda_{k_1},\cdots,\lambda_{k_p})^s$$

$$1\leqslant k_1<k_2<\cdots<k_p\leqslant n$$

的项，而 $X_{k_1\cdots k_p}$ 表示X的一个 p 阶子式，由X的 1 至 p 行和第 $k_1\cdots k_p$ 诸列上的元素组成，$Y_{k_1\cdots k_p}$ 是 Y 的一个 p 阶子式，由 $\overline{Y}$ 的 1 至 p 列及第 $k_1,k_2,\cdots,k_p$ 各行上的元素组成，k_1，$k_2,\cdots,k_p$ 是 $1,2,\cdots,n$ 的任意一个元素个数为 p 的子集，当 k_1，$k_2\cdots,k_p$ 分别是 $1,\cdots,p$ 时，简记为 X_p 和 Y_p。当 $X_pY_p\neq 0$ 时，(3.1.25)右端中的第一项将是 $\Delta_p^{(s)}$的主要部分。

记 $\Delta_{ip}^{(s)}$ 和 $X_p^{(i)}$分别为将$\Delta_p^{(s)}$ 和 X_p 中的第 p 行换为 A^s 和 X的 行后所成的行列式，则

$$\Delta_{ip}^{(s)}=X_p^{(i)}Y_p(\lambda_1,\cdots,\lambda_p)^s+\cdots, \tag{3.1.26}$$

于是当 $X_p\overline{Y}_p\neq 0$ 时，有

$$t_{ip}^{(s)}=\frac{X_p^{(i)}Y_p(\lambda_1,\cdots,\lambda_p)^s+\cdots}{X_p\overline{Y}_p(\lambda_1,\cdots,\lambda_p)^s+\cdots}$$

$$=\frac{X_p^{(i)}}{X_p}+O\left(\frac{\lambda_{p+1}}{\lambda_p}\right)^s, \tag{3.1.27}$$

当 $s\to\infty$ 时，$t_{ip}^{(s)}$ 有极限，而当 $X_p=0$，$Y_p\neq 0$，$X_p^{(i)}\neq 0$ 时，$t_{ip}^{(s)}$将是发散的。

据此，乃有下列关于矩阵的 LR 分解序列收敛的定理。

定理 3.1.2 设A有 LR 分解，A的特征值依模不等，并且A的特征向量所成的矩阵以及该矩阵的逆的各阶首主子式不为零，则A的 LR 分解序列必收敛于某一上三角阵。而极限矩阵的对角元就是矩阵A的特征值。

证明 定理的条件保证了 $t_{ip}^{(s)}(i>p)$的收敛性，但(3.1.27)

右端的极限恰是X的LR分解中的下三角阵的(i,p)元素，即

$$T_s \to T \tag{3.1.28}$$

这里T是X的LR分解$X=TS$中的单位下三角阵

$$\begin{aligned} A_s &= T_{s-1}^{-1} A T_{s-1} \to T^{-1} A T \\ &= S\mathrm{diag}(\lambda_i) S^{-1}. \end{aligned} \tag{3.1.29}$$

上式右端是以$\lambda_1, \lambda_2, \cdots, \lambda_n$为对角的上三角阵。

当$A=A_1$是对称正定阵时，由于X可取为正交阵，有$Y=X^{-1}=X^T$，可得

$$X_{i_1\cdots i_p} = Y_{i_1\cdots i_p}$$

及

$$t_{i,p}^{(s)} = \frac{\sum X_{i_1\cdots i_p}^{(i)} Y_{i_1\cdots i_p} (\lambda_{i_1}, \cdots, \lambda_{i_p})^s}{\sum (X_{i_1\cdots i_p})^2 (\lambda_{i_1}, \cdots, \lambda_{i_p})^s} \tag{3.1.30}$$

这里$X_{i_1\cdots i_p}^{(i)}$表示$X_{i_1\cdots i_p}$中把第p行的元素换为X的第i行$(i>p)$的元素后所成的行列式。将(3.1.30)两端取极限，可见，当$s\to\infty$时，$t_{i,p}^{(s)}$不会遇到发散的情况，记

$$T_s \to T_\infty \tag{3.1.31}$$

此时T_∞未必是T，但仍是单位下三角阵，而由(3.1.31)

$$L_s = T_{s-1}^{-1} T_s \to T_\infty^{-1} T_\infty = I$$

但

$$A_s = T_{s-1}^{-1} A T_{s-1} \to T_\infty^{-1} A_1 T_\infty \tag{3.1.32}$$

$$R_s = L_s^{-1} A_s \to T_\infty^{-1} A_1 T_\infty$$

可见A_s与R_s有同样的极限，由于R_s是上三角阵，故其极限也应是上三角阵，并且该极限阵的对角元应是原矩阵A的特征值。

§2 QR算法及收敛性

2.1 QR算法的基本收敛性

由于矩阵的LR分解中，其单位下三角阵中的元素不能保证依模小于1，甚至无法估计出这些元素模的上界，因此，LR

算法的稳定性便没有保证，实践经验证明，在这种算法下，计算特征值时精度的损失，通常并不是偶然的现象。1961年以来，Francis对此法进行了改进，用酉变换作工具，用QR分解代替LR分解而发展起来的QR方法，在大多数情况下被证明是求一般代数特征值问题的最有效的方法。

令 $A=A_1$ 是非奇异实阵，则A可表示为

$$A_1=Q_1R_1, \tag{3.2.1}$$

其中 Q_1 是正交阵，R_1 是上三角阵，当规定 R_1 的对角元是非负的元素时，此种分解将是唯一的。令

$$A_2=R_1Q_1. \tag{3.2.2}$$

易证 A_2 相似于 A_1，一般有

$$A_1=A,\quad A_s=Q_sR_s,\quad A_{s+1}=R_sQ_s, \tag{3.2.3}$$

$$s=1,2,\cdots.$$

这里每个 Q_s 仍是正交阵，R_s 仍是对角元非负的上三角阵。不难证明

$$A_{s+1}=Q_s^H\cdots Q_1^HA_1Q_1\cdots Q_s, \tag{3.2.4}$$

$$Q_1\cdots Q_sA_{s+1}=A_1Q_1\cdots Q_s. \tag{3.2.5}$$

令

$$Q_1\cdots Q_s=P_s,\quad P_s\cdots P_1=U_s,$$

则

$$A_1^s=P_sU_s. \tag{3.2.6}$$

在讨论A的QR分解序列的收敛性之前，先考察一个例子。

例3.2.1　令

$$A=\begin{pmatrix} e^{i\theta} & 1 \\ & e^{-i\theta} \end{pmatrix},$$

则其QR分解序列可表为

$$A_s=\begin{pmatrix} e^{i\theta} & e^{-2(s-1)i\theta} \\ & e^{-i\theta} \end{pmatrix}.$$

这里，由于(1,2)元素在 $s\to\infty$ 时无极限，所以 A_s 在通常的意义下是不收敛的。然而，对于 A_s 的下三角阵收敛于对角阵的情况，不管 A_s 的严格上三角阵有无极限，只要相应的元素保持有界，且把相应的对角阵当作 A_s 的极限，则 A_s 在实质上将收敛于此对角阵。在本章后面的讨论中，当谈到矩阵列的极限时，我们都把它当作依实质收敛来理解。

下面我们来讨论 A 的 QR 分解序列 $\{A_s\}$ 收敛的充分条件。

定理 3.2.1 设 A 非奇异且其特征值依模不等：

$$|\lambda_1|>|\lambda_2|>\cdots>|\lambda_n|>0 \tag{3.2.7}$$

设其特征向量所成的矩阵为 X，当 X 的逆 Y 有 LR 分解，即

$$X^{-1}=Y=L_yU_y \tag{3.2.8}$$

其 $\{A_s\}$ 将在实质上收敛于依次以 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 为对角元的上三角阵。

证明 由

$$A=X\mathrm{diag}(\lambda_1,\cdots,\lambda_n)Y=X\Lambda Y,$$

$$A^s=X\Lambda^sY,\quad \Lambda=\mathrm{diag}(\lambda_1\cdots\lambda_n), \tag{3.2.9}$$

代入 X 的 QR 分解及 Y 的 LR 分解：

$$X=Q_xR_x,\quad Y=L_yU_y, \tag{3.2.10}$$

则

$$A^s=Q_xR_x\Lambda^sL_yU_y=Q_xR_x(\Lambda^sL_y\Lambda^{-s})(\Lambda^sU_y),$$

易见 $\Lambda^sL_y\Lambda^{-s}$ 仍是单位下三角阵，其 (i,j) 元素 $(j<i)$

$$l_{ij}(\lambda_i/\lambda_j)^s,\quad (j<i) \tag{3.2.11}$$

是趋于零的，这里 l_{ij} 是 L_y 的 (i,j) 元素。记

$$\Lambda^sL_y\Lambda^{-s}=I+E_s,\quad E_s\to 0, \tag{3.2.12}$$

则

$$A^s=Q_xR_x(I+E_s)\Lambda^sU_y$$
$$=Q_x(I+R_xE_sR_x^{-1})R_x\Lambda^sU_y,$$

记

$$D_1=\begin{pmatrix} e^{i\theta_1} & & & \\ & e^{i\theta_2} & & \\ & & \ddots & \\ & & & e^{i\theta_n}\end{pmatrix},\quad \tilde{\Lambda}=\begin{pmatrix} |\lambda_1| & & & \\ & |\lambda_2| & & \\ & & \ddots & \\ & & & |\lambda_n|\end{pmatrix}.$$

这里 D_1是酉对角阵，$\tilde{\Lambda}$ 是对角元非负的对角阵，类似地，记

$$U_y=D_2\tilde{U}_y,\quad I+R_xE_sR_x^{-1}=\tilde{Q}_s\tilde{R}_s. \tag{3.2.13}$$

这里 D_2 是酉对角阵，$\tilde{U}_y$ 是对角元非负的上三角阵，$\tilde{Q}_s$、$\tilde{R}_s$ 是 $I+R_xE_sR_x^{-1}$的 QR 分解，于是

$$\begin{aligned} A^s &= Q_x\tilde{Q}_s\tilde{R}_sR_xD_1^s\tilde{\Lambda}^sD_2\tilde{U}_y \\ &= Q_x\tilde{Q}_s\tilde{R}_sR_xD_1^sD_2\tilde{\Lambda}^s\tilde{U}_y \\ &= Q_x\tilde{Q}_sD_1^sD_2\tilde{R}_s\tilde{R}_x\tilde{\Lambda}^s\tilde{U}_y. \end{aligned}$$

由于对角阵的乘法是可换的，而酉对角阵右乘于对角元非负的的上三角阵，并从左边提出此酉对角阵时，上三角阵中的对角元将保持不变，即仍然为正，故上式中 $\tilde{R}_s$、$\tilde{R}_x$ 仍是对角元为正的上三角阵。由 (3.2.13) 推知 $\tilde{Q}_s$、$\tilde{R}_s$ 将收敛于单位阵，并且 $\tilde{R}_s$ 将收敛于单位阵。由A^s 的 QR 分解的唯一性，应有 $A^s=P_sU_s$ 及

$$P_s=Q_x\tilde{Q}_sD_1^sD_2,\quad U_s=\Lambda_s\tilde{R}_x\tilde{\Lambda}^s\tilde{U}_y, \tag{3.2.14}$$

又由

$$P_s=Q_1\cdots Q_s=P_{s-1}Q_s,$$

有

$$\begin{aligned} Q_s &= P_{s-1}^{-1}P_s=D_2^{-1}D_1^{-s+1}\tilde{Q}_{s-1}^{-1}Q_x^{-1}Q_x\tilde{Q}_sD_1^sD_2 \\ &= D_1^{-s+1}D_2^{-1}\tilde{Q}_{s-1}^{-1}\tilde{Q}_sD_1^sD_2 \\ &= D_1^{-s+1}D_2^{-1}(I+F_s)D_1^sD_2 \\ &= (I+\tilde{F}_s)D_1^{-s+1}D_2^{-1}D_1^sD_2 \\ &= (I+\tilde{F}_s)D_1\to D_1. \end{aligned} \tag{3.2.15}$$

这里 F_s、$\tilde{F}_s$ 是根据 $\tilde{Q}_s$ 收敛于单位阵而导出的收敛于零的阵。类似可证

$$\begin{aligned} R_s &= U_s U_{s-1}^{-1} \\ &= \tilde{\tilde{R}}_s \tilde{R}_x \tilde{\Lambda}^s \tilde{U}_y \tilde{U}_y^{-1} \tilde{\Lambda}^{-s+1} \tilde{R}_x^{-1} \tilde{\tilde{R}}_{s-1}^{-1} \\ &= \tilde{\tilde{R}}_s \tilde{R}_x \tilde{\Lambda} \tilde{R}_x^{-1} \tilde{\tilde{R}}_{s-1}^{-1} \to \tilde{R}_x \tilde{\Lambda} \tilde{R}_x^{-1}. \end{aligned}$$

故

$$A_s = Q_s R_s \to D_1 \tilde{R}_x \tilde{\Lambda} \tilde{R}_x^{-1} = \Lambda T \tag{3.2.16}$$

这里 T 是对角元为 1 的上三角阵。

2.2 等模特征值的情形

当条件 (3.2.7) 不满足时，如果 A 只有线性初等因子，这时 A 的 QR 分解序列的收敛性证明需要稍加修改，得到的结论亦将有一些差别。设

$$|\lambda_1| > \cdots > |\lambda_r| = |\lambda_{r+1}| = \cdots = |\lambda_t| > |\lambda_{t+1}| > \cdots > |\lambda_n| > 0 \tag{3.2.17}$$

并且 Y 也有 LR 分解，此时，因

$$A^s = X\Lambda^s L_y R_y = X(\Lambda^s L_y \Lambda^{-s})(\Lambda^s R_y),$$

其中 $\Lambda^s L_y \Lambda^{-s}$ 的位于区域

$$t \geqslant i > j \geqslant r \tag{3.2.18}$$

中的元素为

$$\begin{aligned} l_{ij}(\lambda_i/\lambda_j)^s &= l_{ij}\left(\frac{|\lambda_i| e^{i\theta_i}}{|\lambda_j| e^{i\theta_j}}\right)^s \\ &= l_{ij} e^{i(\theta_i - \theta_j)s}. \end{aligned} \tag{3.2.19}$$

它们的绝对值保持为 $|l_{ij}|$ $(t \geqslant i > j \geqslant r)$，而不随 $s \to \infty$ 而趋于零，但 $\Lambda^s L_y \Lambda^{-s}$ 的在上述区域之外属于下三角阵的不在对角线上的元素则是趋于零的，于是可写

$$A^s = X(\tilde{L}_s + E_s)(\Lambda^s R_y). \tag{3.2.20}$$

这里 $\tilde{L}_s$ 仍是单位下三角阵，其在 (3.2.18) 所示区域部分的元素相应地取为 $l_{ij}(\lambda_i/\lambda_j)^s$，其余部分与单位阵相同，$E_s$ 则是以 O 为极限的阵。记

$$X\tilde{L}_s=\tilde{Q}_s\tilde{U}_s,$$

则

$$\begin{aligned}
A^s &= \tilde{Q}_s\tilde{U}_s(I+\tilde{L}_s^{-1}E_s)\Lambda^s R_y\\
&= \tilde{Q}_s(I+\tilde{U}_s\tilde{L}_s^{-1}E_s\tilde{U}_s^{-1})\tilde{U}_s\Lambda^s R_y\\
&= \tilde{Q}_s(I+F_s)\tilde{U}_s\Lambda^s R_y\\
&= \tilde{Q}_s\bar{Q}_s\bar{R}_s\tilde{U}_s\Lambda^s R_y,
\end{aligned}\tag{3.2.21}$$

其中 $F_s=\tilde{U}_s\tilde{L}_s^{-1}E_s\tilde{U}_s^{-1}\to O$，$\bar{Q}_s\bar{R}_s$ 是 $I+F_s$ 的 QR 分解，由 $F_s\to O$，可得 $\bar{Q}_s$，$\bar{R}_s\to I$，注意，X 与 $X\tilde{L}_s$ 除第 r 列到第 t 列不同外，其余都是相同的，$X\tilde{L}_s$ 的第 r 列到第 t 列分别是 X 的第 r 列到第 t 列的线性组合，若令

$$X=Q_xR_x$$

为 X 的 QR 分解，则由 $X\tilde{L}_s=\tilde{Q}_s\tilde{U}_s$ 及 $X\tilde{L}_s=Q_xR_x\tilde{L}_s$，由于 $R_x\tilde{L}_s$ 具有形状

$$R_x\tilde{L}_s=\left(\begin{array}{cccccccccc}
\times & \times & \cdots & \times & \triangle & \cdots & \triangle & \times & \cdots & \times\\
 & \times & & \vdots & \vdots & & \vdots & \vdots & & \vdots\\
 & & & \times & \vdots & & \vdots & \vdots & & \vdots\\
 & & & & \triangle & \cdots & \triangle & \vdots & & \\
 & & & & \vdots & & \vdots & \vdots & & \\
 & & & & \triangle & \cdots & \triangle & \times & & \\
 & & & & & & & \times & & \\
 & & & & & & & & \ddots &
\end{array}\right)\begin{array}{l} \\ \\ \\ r\\ \\ t\\ \\ \\ \end{array},$$

其中画“×”的元素与 R_x 中相应的元素相等，画“△”的元素则是 R_x 的第 t 列到第 r 列的元素的线性组合。

设 $R_x\tilde{L}_s$ 的 QR 分解为

$$R_x\tilde{L}_s=Q_\Delta^{(s)}R_\Delta^{(s)},$$

则由(3.2.22)的右端可知，$Q_{\Delta}^{(s)}$除第 t 到 r 行、列为一正交阵外，其余与单位阵相同，而 $R_{\Delta}^{(s)}$在 r 行上方各行和 t 行下方各行与 $R_x\tilde{L}_s$ 的相应的各行一致，比较

$$X\tilde{L}_s = \tilde{Q}_s\tilde{U}_s = Q_x Q_{\Delta}^{(s)} R_{\Delta}^{(s)}, \tag{3.2.22}$$

可见

$$\tilde{Q}_s = Q_x Q_{\Delta}^{(s)},\ \tilde{U}_s = R_{\Delta}^{(s)}.$$

$\tilde{Q}_s$ 与 Q_x 在第 r 列以左第 t 列以右是一样的，$\tilde{Q}_s$ 的第 r 至第 t 列则是 Q_x 的第 r 至第 t 列的组合，据(3.2.21)比较

$$A^s = P_s U_s = \tilde{Q}_s \bar{Q}_s \bar{R}_s \tilde{U}_s \Lambda^s R_y,$$

记 $\Lambda^s = D^s|\Lambda|^s$，其中 D^s 是酉对角阵，把上式 A^s 中的酉对角阵 D^s 往左移，可见

$$P_s = \tilde{Q}_s \bar{Q}_s D^s.$$

据 $A_{s+1} = P_s^H A_1 P_s = (D^H)^s \bar{Q}_s^H \tilde{Q}_s^H A_1 \tilde{Q}_s \bar{Q}_s D_s$ 及 (3.2.22) 可得，$A_{s+1} = (D^H)^s \bar{Q}_s^H \tilde{U}_s \tilde{L}_s^{-1} X^{-1} A_1 X \tilde{L}_s \tilde{U}_s^{-1} \bar{Q}_s D^s$，由于 $\bar{Q}_s \to I$，D_s 是酉对角阵，故知 A_{s+1} 依实质收敛于

$$\tilde{U}_s(\tilde{L}_s^{-1} X^{-1} A_1 X \tilde{L}_s)\tilde{U}_s^{-1}.$$

注意，$X\tilde{L}_s$ 的第 r 列以左和第 t 列以右应是 A_1 的对应于 λ_1，…，λ_{r-1}，λ_{t+1}，…，λ_n 的特征向量，于是 $\tilde{L}_s^{-1} X^{-1} A_1 X \tilde{L}_s$ 有下列形状

$$\tilde{L}_s^{-1} X^{-1} A_1 X \tilde{L}_s = \begin{pmatrix} \lambda_1 & & & & & & * \\ & \ddots & & & & & \\ & & \lambda_{r-1} & & & & \\ & & & \begin{matrix} \times & \cdots & \times \\ \vdots & & \vdots \\ \times & \cdots & \times \end{matrix} & & & \\ & & & & \lambda_{t+1} & & \\ & & & & & \ddots & \\ & & & & & & \lambda_n \end{pmatrix}, \tag{3.2.23}$$

$\widetilde{U}_s$与$\widetilde{U}_s^{-1}$是互逆的上三角阵，分别左乘和右乘于(3.2.23)后，可知A_{s+1}依实质收敛于形如(3.2.23)的矩阵。

具有等模特征值的矩阵的一种特殊情况是，矩阵A只具有单重实特征值和单重共轭复特征值。此时，A的QR分解序列将依实质收敛于前述的拟上角阵。当矩阵有多个等模特征值分布在以原点为圆心的同心圆周上时，特征值的确定将不如前面所述的情况直接，这种困难将在后面予以克服。

定理3.2.1中要求$X^{-1}=Y$有LR分解这个条件也可以去掉，但在这种情况下，QR分解的序列的极限矩阵的对角线上的元素可能不再按依模递减的次序来排列。有关的证明可参看J. H. Wilkinson的书的第8章§31。

§3 带原点位移的QR算法

3.1 带位移的QR分解

当A的特征值依模不等时，如果Y阵又有LR分解，则A的QR分解序列将收敛于以λ_1，λ_2，…，λ_n为对角元的上三角阵，A_s的下三角部分的元素将以$(\lambda_i/\lambda_j)^s(i>j)$的敛速收敛于零。由于$A$的特征值与$A-pI$的特征值之差是常数$p$，而$A-pI$的QR分解序列的下三角部分的元素将以$(\lambda_i-p)^s/(\lambda_j-p)^s$的敛速收敛于零，当$p$相对地更接近$\lambda_n$时，$(\lambda_n-p)^s/(\lambda_{n-1}-p)^s$收敛于零的速度就特别快。如果我们有一个确定$A$的依模最小的特征值$\lambda_n$的一个近似方法，即若

$$k_s\to\lambda_n \qquad (s\to\infty) \tag{3.3.1}$$

则QR分解过程必将得到加速。作

$$A_s-k_sI=Q_sR_s,\quad A_{s+1}=R_sQ_s+k_sI \tag{3.3.2}$$

$$s=1,2,\cdots$$

称 $\{A_s\}$ 为A的带恢复的以 k_s 为位移量的 QR分解，可以证明

$$A_{s+1}=R_sQ_s+k_sI=Q_s^HA_sQ_s,$$

$$A_{s+1}=Q_s^H\cdots Q_1^HA_1Q_1\cdots Q_s, \tag{3.3.3}$$

$$Q_1\cdots Q_sA_{s+1}=A_1Q_1\cdots Q_s, \tag{3.3.4}$$

$$Q_1\cdots Q_{s-1}(A_s-k_sI)R_{s-1}\cdots R_1$$
$$=(A_1-k_sI)(A_1-k_{s-1}I)\cdots(A_1-k_1I). \tag{3.3.5}$$

又当适当选择 k_s 时，则在 A_s 中，其第 n 行将迅速出现收敛的情况，当 $a_{nj}^{(s)}(j<n)$ 在计算精度要求下趋近于零时，便可将 $a_{nn}^{(s)}$ 作为A的一个近似特征值输出，A的其余特征值可在A_s 的降维后的矩阵中去寻找。

当A的依模最小的特征值是实的且是单重的时，可取 $k_s=a_{nn}^{(s-1)}$。$(s=1,2,\cdots)$。

当A的依模最小的特征值是共轭复数时，可取 k_s 为A_{s-1}的右下角的二阶矩阵：

$$\begin{vmatrix} a_{n-1\,n-1}^{(s-1)} & a_{n-1\,n}^{(s-1)} \\ a_{n\,n-1}^{(s-1)} & a_{nn}^{(s-1)} \end{vmatrix}$$

的特征值，这时 k_s 可能是复数。为了在实际上使用复位移而又能避免复数运算，需要进一步研究 QR 分解方法。

3.2 共轭复位移

设 k_1，k_2 是共轭的两个复数，作

$$A_1-k_1I=Q_1R_1,\quad A_2=R_1Q_1+k_1I, \tag{3.3.6}$$

$$A_2-k_2I=Q_2R_2,\quad A_3=R_2Q_2+k_2I,$$

则 A_3是 A_1连续经 k_1、k_2 两次复位移后所得的矩阵，当 A_1是实阵时，虽然 Q_1、R_1、Q_2、R_2 都是复阵，但因

$$A_3=(Q_1Q_2)^HA_1(Q_1Q_2). \tag{3.3.7}$$

若令

$$Q=Q_1Q_2,\quad R=R_2R_1, \tag{3.3.8}$$

则有

$$QR = Q_1Q_2R_2R_1 = (A_1 - k_2I)(A_1 - k_1I)$$

$$= A_1^2 - (k_1 + k_2)A_1 + k_1k_2I. \qquad (3.3.9)$$

上式的右端是实阵，它的QR分解也应是实阵，因而Q、R也是实的，由（3.3.7），从而A_3也是实的。

3.3 Hessenberg阵

为了简化计算，我们证明下述定理。

定理 3.3.1 设A为实的非奇异阵，Q是正交阵，B是次对角元素为正数的实的上Hessenberg阵，则在等式

$$AQ = QB \qquad (3.3.10)$$

中，当Q的第1列给定时，Q、B将由(3.3.10)唯一确定。

证明 设$Q = (q_1, \cdots q_n)$是正交阵，其中q_1是已知的，

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ & b_{32} & \cdots & b_{3n} \\ & \ddots & \ddots & \vdots \\ & & b_{nn-1} & b_{nn} \end{pmatrix}, \qquad (3.3.11)$$

比较(3.3.10)两端的第一列，得

$$Aq_1 = b_{11}q_1 + b_{21}q_2, \qquad (3.3.12)$$

$$b_{11} = q_1^TAq_1, \quad b_{21} = \|Aq_1 - b_{11}q_1\| > 0, \qquad (3.3.13)$$

$$q_2 = (Aq_1 - b_{11}q_1)/b_{21}, \quad b_{21} \neq 0. \qquad (3.3.14)$$

继续比较(3.3.10)两端的第2，3，…，n各列，并仿此讨论，知B、Q均被唯一确定：

$$b_{ik} = q_i^T A q_k (i = k, \cdots, n),$$

$$b_{k+1,k} = \|Aq_k - \sum_{i=1}^{k} b_{ik}q_i\| > 0, \qquad (3.3.15)$$

$$q_{k+1} = (Aq_k - \sum_{i=1}^{k} b_{ik}q_i)/b_{k+1,k},$$

$$k = 1, 2, \cdots, n-1.$$

注意，该定理把诸$b_{k+1,k}\neq 0(k=2,\cdots,n-1)$作为条件，但此条件是否成立应在(3.3.15)的递推计算中才能发现，一旦在某一步当$b_{k+1,k}=0$时，q_{k+1}就不能再由(3.3.15)确定了，此时，可把$\boldsymbol{q}_{k+1}$取作与$\boldsymbol{q}_i(i=1,\cdots,k)$正交的任意一个单位向量，递推仍可继续下去，但$Q$、$B$的唯一性就不再能保证了。

定理 3.3.2 设A是非奇异的上 Hessenberg 阵且有非零次对角元，则A的 LR 和 QR 分解序列$\{A_s\}$仍为上 Hessenberg 阵，且有非零次对角元。

证明 设

$$A=LR \quad 或 \quad A=QU, \tag{3.3.16}$$

比较两端第1列，有

$$\boldsymbol{a}_1=r_{11}\boldsymbol{l}_1 \quad 或 \quad \boldsymbol{a}_1=u_{11}\boldsymbol{q}_1. \tag{3.3.17}$$

由于$\boldsymbol{a}_1$具有形状：

$$\boldsymbol{a}_1^T=(x,x,0,\cdots,0), \tag{3.3.18}$$

可见$\boldsymbol{l}_1^T$、$\boldsymbol{q}_1^T$亦具有(3.3.18)的形状，依次比较(3.3.16)的其余各列，可见L、Q亦均为上 Hessenberg 阵，且有非零次对角元，再由

$$\boldsymbol{A}_2=\boldsymbol{RL} \quad 或 \quad \tilde{A}_2=UQ$$

及$\boldsymbol{R}$、$\boldsymbol{U}$是上三角阵，可知两个A_2仍是上 Hessenberg 阵。同理$\boldsymbol{A}_3,\cdots,\boldsymbol{A}_s,\cdots$均是上 Hessenberg 阵。

由于任意矩阵均可经由 Householder 相似变换化为上 Hessenberg 阵，当其次对角线上有零元时，此种矩阵的特征值问题可以化为若干个低阶的次对角元非零的上Hessenberg阵的特征值问题。因此，不失一般性，今后我们将只讨论次对角元非零的上 Hessenberg 阵。它又被称为不可约的上 Hessenberg 阵。由于 Hessenberg 阵的 LR 或 QR 分解要比一般的满阵的分解在计算上省得多，且又由于定理 3.3.2 的理由，通常计算矩

阵的特征值时，总是先把给定矩阵化为上 Hessenberg 阵.在下面的讨论中,我们将始终假定 A_1 是不可约的上 Hessenberg 阵.

把定理 3.3.1 用到带两步共轭复位移的 QR 分解，由

$$\begin{cases} A_3 = Q^H A_1 Q, \quad Q = Q_1 Q_2, \\ A_1 Q = Q A_3, \end{cases} \tag{3.3.19}$$

可见,

$$A_1 \tilde{Q} = \tag{3.3.20}$$

中，当取 $\tilde{Q}$ 的第 1 列为 Q 的第 1 列时，则由上式确定的 $\tilde{Q}$、B 便应分别是 Q 和 A_3。

由（3.3.9）：$QR = A_1^2 - (k_1 + k_2) A_1 + k_1 k_2 I$，两端取第 1 列，得

$$r_{11} q_1 = (A_1 - k_1 I)(A_1 - k_2 I) e_1 \tag{3.3.21}$$

上式右端的列向量，在 A_1 是 Hessenberg 阵的假定下只有三个非零分量，即前三个分量，分别记为 x_1, y_1, z_1 则

$$\begin{aligned} x_1 &= a_{11}^2 + a_{12} a_{21} - a_{11}(k_1 + k_2) + k_1 k_2, \\ y_1 &= a_{21}(a_{11} + a_{22} - k_1 - k_2), \\ z_1 &= a_{21} a_{32} \end{aligned} \tag{3.3.22}$$

于是

$$\begin{aligned} q_1^T &= \pm (x_1, y_1, z_1, 0, \cdots, 0)/k, \\ k &= (x_1^2 + y_1^2 + z_1^2)^{1/2}. \end{aligned} \tag{3.3.23}$$

3.4 Francise 方法

有了 q_1 之后，本来就可按照（3.3.12）~（3.3.15）的公式计算(3.3.19)中的 Q 和 A_3 了。但是，按照这个过程去进行计算时，如果遇到上 Hessenberg 阵 A_3 的次对角元依模很小的情况，依(3.3.15)计算出的 $b_{k+1\,k}$ 即 $a_{k+1\,k}^{(3)}$ 和 q_{k+1} 的精度便很低，这将直接影响 Q 阵的正交性。因此，从计算稳定性看，这种算法是不可取的。Francise 对此做了改进，其计算过程可以归结为：求 $(n-1)$ 个 Householder 阵 $P_1, \cdots, P_{n-1}$，使 P_1 的第 1 列为 q_1，使 $P_1 \cdots P_k$，$(k = 1, 2, \cdots, n-1)$ 的第 1 列也保持为 q_1，但

$$P_{n-1}^H \cdots P_1^H A_1 P_1 \cdots P_{n-1} \tag{3.3.24}$$

应是上 Hessenberg 阵。下面我们就来描述这些矩阵的构造方法。设

$$P_1 = I - 2\mathbf{w}_1 \mathbf{w}_1^T, \tag{3.3.25}$$

由于P_1的第 1 列应是q_1，而 q_1 只有前三个分量不为零，故 $\mathbf{w}_1$ 也是这样，令

$$\mathbf{w}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|, \quad \mathbf{v}_1^T = (1, \varphi_1, \varphi_2, 0, \cdots 0), \tag{3.3.26}$$

则

$$\|\mathbf{v}_1\|^2 = 1 + \varphi_1^2 + \varphi_2^2, \quad P_1 = I - 2\mathbf{v}_1 \mathbf{v}_1^T / \|\mathbf{v}_1\|^2,$$

$$(P_1 \mathbf{e}_1)^T = \left(1 - \frac{2}{\|\mathbf{v}_1\|^2}, \ -\frac{2\varphi_1}{\|\mathbf{v}_1\|^2}, \ -\frac{2\varphi_2}{\|\mathbf{v}_1\|^2}, 0, \cdots, 0\right)$$

$$= \pm (x_1, y_1, z_1, 0, \cdots, 0)/k, \tag{3.3.27}$$

故

$$\frac{2}{\|\mathbf{v}_1\|^2} = \frac{k \pm x_1}{k}. \tag{3.3.28}$$

为使上式中的分子不致相消，令

$$k = \operatorname{sign}(x_1)(x_1^2 + y_1^2 + z_1^2)^{1/2}, \tag{3.3.29}$$

得

$$\varphi_1 = y_1/(k + x_1), \qquad \varphi_2 = z_1/(k + x_1), \tag{3.3.30}$$

$$\frac{2}{\|\mathbf{v}_1\|^2} = (k + x_1)/k; \tag{3.3.31}$$

于是P_1具有下列形状

$$P_1 = \left(\begin{array}{ccc:c} \times & \times & \times & \\ \times & \times & \times & 0 \\ \times & \times & \times & \\ \times & \times & \times & \\ \hdashline & 0 & & I_* \end{array}\right). \tag{3.3.32}$$

这是一个三阶 Householder 阵与一个单位阵的直接和。由于 $P_1 = P_1^H$，记 $B_1 = P_1 A P_1$，则 B_1 具有下列形状：

$$
B_1 = \begin{pmatrix}
\times & \times & \times & \cdots & \cdots & \cdots & \times \\
\times & \times & \times & & & & \times \\
\times & \times & \times & & & & \vdots \\
\times & \times & \times & \times & & & \vdots \\
 & & & \times & \times & & \vdots \\
 & & & & & \ddots & \vdots \\
 & & & & & \times & \times
\end{pmatrix}, \qquad (3.3.33)
$$

其第 4 行 4 列以后仍保持为上 Hessenberg阵的形状。取P_2 为：

$$
P_2 = \left(\begin{array}{c:c}
1 & 0 \\
\hdashline
0 & U_2
\end{array}\right), \qquad (3.3.34)
$$

其中U_2为 $(n-1)$ 阶 Householder 阵，它把

$$
B_1 = \left(\begin{array}{c:c}
b_{11}^{(1)} & B_{12}^{(1)T} \\
\hdashline
b_{21}^{(1)} & \\
b_{31}^{(1)} & \\
b_{41}^{(1)} & \\
0 & B_{22}^{(1)} \\
\vdots & \\
0 &
\end{array}\right) = \left(\begin{array}{c:c}
b_{11}^{(1)} & B_{12}^{(1)T} \\
\hdashline
B_{21}^{(1)} & B_{22}^{(1)}
\end{array}\right) \qquad (3.3.35)
$$

中的 $B_{21}^{(1)}$ 变形为

$$U_2 B_{21}^{(1)} = \begin{pmatrix} \bar{b}_{21}^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.3.36}$$

的形式。此时，U_2 实际上仍是一个三阶Householder阵与一个单位阵的直接和，而 $B_2 = P_2 B_1 P_2$ 则具有下列形状：

$$B_2 = P_2 B_1 P_2 = \left(\begin{array}{c:ccc} \times & \times & \cdots & \times \\ \hdashline \times & & & \\ 0 & & B_{22}^{(2)} & \\ \vdots & & & \\ 0 & & & \end{array}\right). \tag{3.3.37}$$

它的第 1 列已经化为上Hessenberg型，$B_{22}^{(2)}$ 中除左上角在前三列含有一个 4×3 的 矩 阵 外，其 第 4 行 4 列 以后仍保持为上Hessenberg 型，$B_{22}^{(2)}$的形状与B_1的形状是一样的，因此，此后的计算便可组织在一个循环之中。每个 P_i 除P_1 外，其任务是把B_{i-1}的第 i 列化成上 Hessenberg 型，而每个 P_i 的计算将归结为一个 U_i 的计算，其计算可仿（3.3.25）~（3.3.31）的 计算进行。

这样，我们就有了一个计算矩阵A的带两步复共轭位移的QR 分 解的实的计算方法，其中分解阵Q具有较好的正交性。

另外，前曾指出，当矩阵有等模特征值时，QR分解序列的极限将不再具有拟上三角阵的形式，极限阵中将出现阶数大于2的对角块，这种矩阵的特征值难于处理。在有了带位移特别是带两步共轭复位移的算法以后，矩阵有等模特征值的困难就不再严峻了。如 图 3.1，设A为某实阵，其等模特征值分布在以原点为中心，以 r 为半径的某圆上且对称地出现在x-轴的两侧。

当产生位移时，新原点如靠近某一特征值，必也靠近与其共轭的特征值，但同时却远离了其余的特征值，使原来有几对等模特征值的问题，在位移以后，化为只有一对最小的等模特征值问题。可见，使用带位移的QR算法，不仅能加速收敛过程，而且其适应能力也很强。

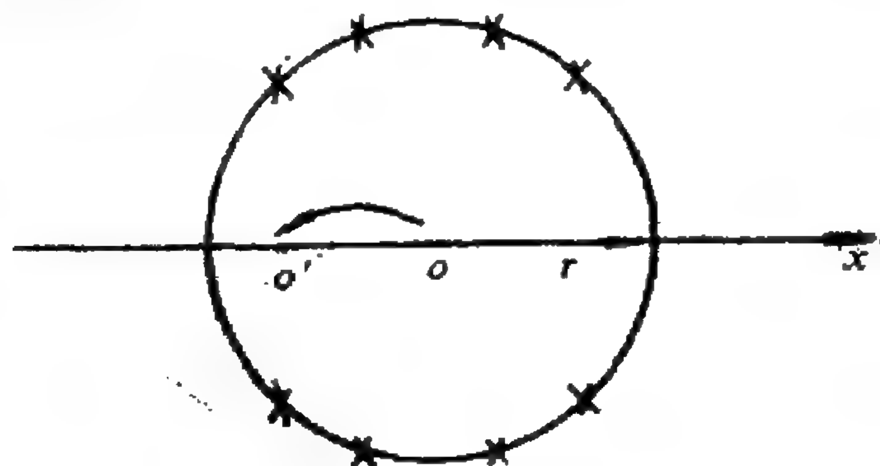


图 3.1 等模特征值

§4 QR 方法的细节

4.1 QR_2子程序

用 QR_2 代表由 Hessenberg 矩阵A计算它的带两步位移量（实或共轭复数）的QR分解的实际计算程序，这个程序将是整个 QR 算法的核心子程序。作为其主要输入的形式参数包括：矩阵的阶数 n，矩阵A和两个以实数形式给出的位移信息

$$\sigma = k_1 + k_2, \quad \rho = k_1 k_2, \tag{3.4.1}$$

其输出则是（3.3.7）中给出的 A_3，它将取代A而占有A原有的存贮单元。

考虑到 $\{A_s\}$ 始终保持着上 Hessenberg 阵的形式，其次对角元在工作精度内可能化为零。出现这种情况时，应分段地把

A_1 的特征值问题化为低阶阵的特征值问题，这样可以大大地节省计算工作量。为了实现这一点，可以从第 n 行往上数，把合于

$$|a_{i,i-1}| < \varepsilon \qquad (i = n,\ n-1, \cdots, 1) \tag{3.4.2}$$

的最大足码记录下来，设为n_p，QR分解序列计算便只在第n_p行（列）到第 n 行（列）之间进行。

在实现

$$A_3 = B_{n-1} = P_{n-1} \cdots P_1 A_1 P_1 \cdots P_{n-1}$$

的计算时，可按递推公式

$$B_j = P_j \quad B_{j-1} P_j \ ,\ j = 1, \cdots, n-1$$

$$B_0 = A_1$$

去组织循环，由

$$\begin{cases} P_j = I - \boldsymbol{v}_j (\alpha \boldsymbol{v}_j^T), \\ \alpha = \left(1 + \dfrac{x_1}{k}\right), \\ k = \text{sign}(x_1)(x_1^2 + y_1^2 + z_1^2)^{1/2}, \\ \boldsymbol{v}_j = \boldsymbol{e}_j + \varphi_1 \boldsymbol{e}_{j+1} + \varphi_2 \boldsymbol{e}_{j+2}, \\ \varphi_1 = y_1/(x_1 + k),\ \ \varphi_2 = z_1/(x_1 + k); \end{cases} \tag{3.4.3}$$

则当 x_1, y_1, z_1 给定时，由 A_1 计算 A_3 的过程就完全可以进行了。令

$$B_j = B'_{j-1} P_j \ , \tag{3.4.4}$$

$$B'_{j-1} = P_j B_{j-1} = (I - \boldsymbol{v}_j (\alpha \boldsymbol{v}_j^T)) B_{j-1}, \tag{3.4.5}$$

$$\boldsymbol{\eta} = \alpha \boldsymbol{v}_j^T B_{j-1}; \tag{3.4.6}$$

则

$$B'_{j-1} = B_{j-1} - \boldsymbol{v}_j \boldsymbol{\eta}^T. \tag{3.4.7}$$

于是

$$\begin{aligned} B_j &= B'_{j-1} - \boldsymbol{\xi} \boldsymbol{v}_j^T, \\ \boldsymbol{\xi} &= \alpha B'_{j-1} \boldsymbol{v}_j \ . \end{aligned} \tag{3.4.8}$$

计算每个 P_j 所需的输入信息 x_1, y_1, z_1 只有一点不同，即 P_1 所需的 x_1, y_1, z_1 是由

$$x_1 = a_{11}^2 + a_{12}a_{21} - \sigma a_{11} + \rho,$$

$$y_1 = a_{21}a_{11} + a_{22}a_{21} - \sigma a_{21},$$

$$z_1 = a_{32}a_{21}$$

计算出来的，(参看 (3.3.22))，而 $P_2, \cdots, P_{n-1}$ 所需的 x_1, y_1, z_1 则应分别由 $B_1, \cdots, B_{n-2}$ 中的元素提供，即

$$\begin{aligned} x_1 &= b_{j+1,j}^{(j)}, \\ y_1 &= b_{j+2,j}^{(j)}, \quad j = 1, 2, \cdots, n-2, \\ z_1 &= b_{j+3,j}^{(j)}, \quad j+3 > n \text{ 时 } z_1 = 0. \end{aligned} \tag{3.4.9}$$

于是有关 $B_1, B_2, \cdots, B_{n-1}$ 的计算几乎可以在同一循环下统一地进行处理了。这样编程的结果，从 A_1 算出 A_3 大约只需 $5n^2$ 次乘法，而在降阶的情况下还会更省一些。

4.2 位移的选择

使用带位移的 QR 分解技术，目的在于加速 QR 分解序列的收敛进程。然而，在当过程显露出某种收敛趋势时，这种技术才是比较有把握的。往往有一些简单的措施，虽不曾在理论上证明是最好的，但却能帮助我们去"捕捉"这种趋势。

把矩阵的右下角的二阶矩阵的特征值 μ_1 和 μ_2 计算并存放起来，称之为老的近似特征值。当 μ_1、μ_2 是复数时，应分别把它们的实部和虚部存放起来。然后，用不带位移的两步 QR 分解法求 A_3 并也求出新的 μ_1 和 μ_2。如果把老的 μ_1、μ_2 和新的 μ_1、μ_2 分别记为 $\mu_1^{(s-1)}$, $\mu_2^{(s-1)}$ 和 $\mu_1^{(s)}$, $\mu_2^{(s)}$，并约定当两个 μ 均为实数时，把数值较小的那个放到 μ_1，另一个放到 μ_2，当两个 μ 是共轭复数时，把有负虚部的那个当作 μ_1 而把有正虚部的当作 μ_2。当

$$|\mu_i^{(s)}-\mu_i^{(s-1)}|>\frac{1}{2}|\mu_i^{(s)}|,\quad (i=1,2) \tag{3.4.10}$$

时，即认为过程还未显露收敛的趋势。这时可令

$$k_i^{(s)}=k_i^{(s-1)}\ (i=1,2). \tag{3.4.11}$$

这表示强令新老位移一致，即仍用原来的位移，而 $k_i^{(0)}=0$，$(i=1,2)$，即在开始时的位移量是取为零的，这里用 k_1, k_2 作存放位移的单元，而 $k_i^{(s)}(i=1,2)$ 是其流动值。当

$$|\mu_i^{(s)}-\mu_i^{(s-1)}|\leqslant\frac{1}{2}|\mu_i^{(s)}|,\quad (i=1,2) \tag{3.4.12}$$

时可取

$$k_i^{(s)}=\mu_i^{(s)}\ (i=1,2), \tag{3.4.13}$$

即用新算出的近似特征值作为位移量，当（3.4.10）只有一个成立时，由于使（3.4.12）成立的那个 $\mu_i^{(s)}$（$i=1$ 或 2），可能是实数也可能是复数，当其是实数时，由于另一个 μ 应当相弃，故应取

$$k_1^{(s)}=k_2^{(s)}=\mu_i^{(s)}\quad (i=1\text{ 或 }2),$$

从而

$$\sigma=2\mu_i^{(s)},\quad \rho=(\mu_i^{(s)})^2.$$

当 $\mu_i^{(s)}$ 是复数时，可取

$$k_1^{(s)}=k_2^{(s)}=\mathrm{Re}(\mu_i^{(s)}),$$

得

$$\sigma=2\mathrm{Re}(\mu_i^{(s)}),\quad \rho=(\mathrm{Re}(\mu_i^{(s)}))^2. \tag{3.4.14}$$

（3.4.14）对 $\mu_i^{(s)}$（$i=1$ 或 2）是实数或复数都是通用的。因此，当（3.4.12）中只有一个成立时，我们就用（3.4.14）去确定新位移的信息。（3.4.10）中的 $\frac{1}{2}$ 因子取得稍大或稍小一点是无关紧要的。

当矩阵有等模特征值时，如等模特征值个数 r 大于 2，条件（3.4.12）可能经多次试验都不满足，正如 QR 方法收敛性证明中所作的分析那样，A_s 中将有一个 r 阶的对角块，其元素是摆动而不收敛的，当所论特征值的模最小时，这个对角块将出现在 A_s 的右下角。为了克服这种情况，可在程序中设置一个记录使（3.4.12）不成立的计数器，当使（3.4.12）不成立的次数超过某一规定的正整数 T_1 时，就把矩阵归结为有等模特征值且又是个数大于 2 的情况。由于矩阵的行列式的绝对值等于矩阵的各特征值的模的乘积。对于等模情况，这个模应等于 $|\det(C)|^{1/r}$。因此，位移量可取为

$$k_1 = k_2 = \pm|\det(C)|^{1/r}. \tag{3.4.15}$$

这里 C 是 A_s 的由第 n_p 行（列）到第 n 行（列）组成的矩阵，r 是它的阶，上式中“±”号的选择可使之与 $\mathrm{trace}(C)$ 同号。T_1 的大小建议取为

$$T_1 = \min(n - n_p + 1, 10).$$

4.3 工作精度与矩阵的预处理

通常可以用矩阵 A 的欧氏范数 $\|A\|_E$ 作为 A 的特征值的上界，如果机器字长为 t，则此时合理的工作精度应取为

$$\varepsilon = 2^{-t}\|A\|_E,\ \|A\|_E^2 = \sum_{i,j} a_{ij}^2. \tag{3.4.16}$$

但在一些实际问题中，$\|A\|_E$ 可能比 A 的依模最大的特征值还大许多，这时，按（3.4.16）来确定 ε 就不合实际了。这时，应利用对角相似变换把 $\|A\|_E$ 适当减小，这个工作叫做对矩阵进行平衡化，其要求是使矩阵的各行各列中依模最大的元素，不要在数量级上有太大的差别。

例 3.4.1 设

$$A=\begin{pmatrix}1 & 100 & & \\ 1 & 1 & 1 & \\ & 100 & 1 & 100 \\ & & 1 & \end{pmatrix},$$

于是 $\|A\|_E=\sqrt{30007}=173.2$，如令 $D=\mathrm{diag}\ (10, 1, 10, 1)$ 则有

$$\tilde{A}=D^{-1}AD=\begin{pmatrix}1 & 10 & & \\ 10 & 1 & 10 & \\ & 10 & 1 & 10 \\ & & 10 & 1\end{pmatrix}.$$

从而 $\|\tilde{A}\|_E=\sqrt{604}=24.5$，矩阵的实际依模最大的特征值为17.1。

4.4 计算次序

用 QR 方法求实矩阵的全部工作内容(包括预处理)如下：

(i) 利用对角相似变换使 $\|\mathbf{A}\|_E$ 适当减小；

(ii) 计算 $\|\mathbf{A}\|_E$，trace(A) 及 $\varepsilon=2^{-t}\|\mathbf{A}\|_E$；

(iii) 用正交相似变换使A化为上 Hessenberg 阵。

(iv) 计算右下角的二阶矩阵的特征值 μ_1 及 μ_2，并给定初始位移量 $k_1=k_2=0$；

(v) 求 Hessenberg 阵的次对角元为零元的位置，确定 n_p，当 $n_p\geqslant n-1$ 时输出A的一个或两个特征值，并使矩阵的阶数减1或2；

(vi) 调用一次 QR_2 程序；

(vii) 重新计算 μ_1 及 μ_2，并依 (3.4.10)～(3.4.14) 重新修改位移量，必要时依 (3.4.15) 确定位移量；

(viii) 重复 (v)～(vii) 的循环，直到算出全部的特征值为止。

例 3.4.2 设

$$A_1=\begin{pmatrix}5.0 & -2.0 & -5.0 & -1.0\\ 1.0 & 0.0 & -3.0 & 2.0\\ & 2.0 & 2.0 & -3.0\\ & & 1.0 & -2.0\end{pmatrix}$$

算得 ${}_1^{(1)}=-1$， ${}_2^{(1)}=1$，开始时置 $\sigma_1=\rho_1=0$，调用一次 QR_2，得

$$A_3=\begin{pmatrix}4.11828 & 3.46451 & -5.39263 & 1.51084\\ -0.29992 & 0.68721 & -3.30064 & -4.23760\\ & 0.93016 & 1.11718 & -0.71200\\ & & 0.22015 & -0.92266\end{pmatrix}.$$

由 A_3 算得 $\mu_1^{(3)}=-0.84268$，$\mu_2^{(3)}=1.03720$，此时条件(3.4.12)均满足，应取 $\sigma_3=\mu_1^{(3)}+\mu_2^{(3)}=0.19452$，$\rho_3=\mu_1^{(3)}\mu_2^{(3)}=-0.87403$。经调用 QR_2，得

$$A_5=\begin{pmatrix}3.88102 & 5.50657 & -2.83249 & -9.8166\\ 0.08301 & 2.29698 & -3.41894 & 3.17136\\ & 1.65698 & -0.18116 & 2.41387\\ & & 0.01060 & -0.99684\end{pmatrix}.$$

由 A_5 得 $\mu_1^{(5)}=-1.02709$，$\mu_2^{(5)}=-0.15091$。此时条件(3.4.12)只对 $\mu_1^{(5)}$ 成立，应取 $\sigma_5=2\mu_1^{(5)}=-2.05418$，$\rho_5=(\mu_1^{(5)})^2=1.05491$，

再调 QR_2 得

$$A_7=\begin{pmatrix}4.01900 & 0.73616 & -6.17208 & 0.85558\\ -0.02762 & -0.19762 & -3.26500 & -3.83089\\ & 1.68124 & 2.17861 & 1.23705\\ & & -\varepsilon & -1.00000\end{pmatrix}.$$

由 A_7 得 $\mu_1^{(7)}=-1$，$\mu_2^{(7)}=2.17861$. 此时条件（3.4.12）也只对 $\mu_1^{(7)}$ 成立，应取 $\sigma_1=2\mu_1^{(7)}=-2$，$\rho_1=(\mu_1^{(7)})^2=1$，得

$$A_9=\begin{pmatrix} 4.01231 & 5.45684 & -3.02039 & -0.83340 \\ -0.00903 & 2.11330 & -3.33301 & 3.23455 \\ & 1.57244 & 0.12561 & 2.40441 \\ & & -0. & -1.00000 \end{pmatrix}.$$

此时（4.3）元素已经在机器精度内化为零，一个特征值 $\lambda_4=-1$，已经可以输出，去掉第 4 行 4 列后，仍取初始位移信息 $\sigma_7=\rho_7=0$，计算在第 1 至 3 行 3 列进行，算得

$$\mu_1^{(9)}=1.119455-2.063361i,$$

$$\mu_2^{(9)}=1.119455+2.063361i,$$

调用 QR_2 后得

$$A_{11}=\left(\begin{array}{ccc:c} 3.99618 & 4.15400 & 4.64849 & * \\ 0.99182 & 0.37161 & 1.17200 & * \\ & -3.74868 & 1.63220 & * \\ \hdashline & & & -1 \end{array}\right).$$

由 A_{11} 得 $\mu_1^{(11)}=1001905-1.99905i$, $\mu_2^{(11)}=1.001905+1.99905i$. 此时条件（3.4.12）对两个 μ 都满足，应取 $\sigma_{11}=\mu_1^{(11)}+\mu_2^{(11)}=2.00381$，$\rho_{11}=\mu_1^{(11)}\mu_2^{(11)}=5.0000$. 再算得

$$A_{13}=\left(\begin{array}{ccc:c} 4.00000 & 5.04835 & -3.65643 & * \\ 0 & 1.87849 & -3.59100 & * \\ & 1.32902 & 0.12106 & * \\ \hdashline & & & -1.0 \end{array}\right).$$

由 A_{13} 得 $\mu_1^{(13)}=1.0-2.0i$，$\mu_2^{(13)}=1.0+2.0i$，于是求得全部特征值为：$\lambda_1=4$，$\lambda_{2,3}=1.0\mp2.0i$，$\lambda_4=-1$.

§5 实对称阵的 QR 算法

5.1 QR 算法与对称性

把 QR 方法应用于对称矩阵，理应得出更进一步的结果。我们首先指出，对称性在 QR 变换下仍是保持的。设 $A = A_1$ 是实对称阵，A_2 是 A_1 经一次QR 分解后得到的矩阵：

$$A_1 = QR, \quad A_2 = RQ, \tag{3.5.1}$$

由于 $A_2 = Q^T A_1 Q$ 及 A_1 的对称性可知，A_2 也是对称的。

根据这个性质，当把 QR 方法应用于对称阵时，矩阵 A 及 A_s 均可只存贮其上三角阵部分，这样存贮量与计算量可大致节约一半。

然而，LR 方法却不具有这种性质。但若把LR分解修改成 Cholesky 分解，却仍能保持对称性。即令

$$A_1 = LL^T, \quad A_2 = L^T L, \tag{3.5.2}$$

则

$$A_2 = L^T A_1 L^{-T}. \tag{3.5.3}$$

可见，A_2 仍是对称的。但在此时，不仅要求A_1 是对称的，而且要求 A_1 是正定的，否则（3.5.2）的分解中将遇到复数运算。这是QR方法应用于对称阵时所不具有而为LR 方法所具有的缺点。

把QR方法与 Cholesky 方法加以比较，可进一步看出二者的差别和联系。由于

$$A_1^s = (Q_1 \cdots Q_s)(R_s \cdots R_1) \tag{3.5.4}$$

及

$$\begin{aligned} A_1^{2s} &= (A_1^s)^T A_1^s = (R_s \cdots R_1)^T (Q_1 \cdots Q_s)^T (Q_1 \cdots Q_s)(R_1 \cdots R_s) \\ &= (R_s \cdots R_1)^T (R_s \cdots R_1), \end{aligned} \tag{3.5.5}$$

而由 A_1 的 Cholesky 分解，有

$$\begin{cases}\widetilde{A}_s = L_s L_s^T, \quad \widetilde{A}_{s+1} = L_s^T L_s, \quad s = 1, 2, \cdots, \\ \widetilde{A}_1 = A_1\end{cases} \tag{3.5.6}$$

可得

$$A_1^{2s} = (L_1 \cdots L_{2s})(L_1 \cdots L_{2s})^T \tag{3.5.7}$$

和

$$A_{s+1} = \widetilde{A}_{2s+1}. \tag{3.5.8}$$

可见，A_1 的 Cholesky 分解序列的第 $2s+1$ 个就是 A_1 的QR分解序列的第 $s+1$ 个。由于实对称阵的 QR 分解必是实的，于是（3.5.8）又是任意实对称阵的奇数次 Cholesky 分解必是实的一个证明。

由于 QR 分解的序列将不改变矩阵的2-范数和欧氏范数，矩阵序列中各元素显然是有界的，而在 Cholesky 分解中，当矩阵近于奇异时，其三角阵因子中的元素的上界，一般是无法估计的，所以从算法稳定性来看，这又是 QR 方法的另一优点。

5.2 QR 算法的三阶收敛性

设 $A = A_1$ 是对称阵，$\{A_s\}$ 是由 A_1 产生的QR分解序列，记

$$A_s = \begin{bmatrix} A_{11}^{(s)} & A_{12}^{(s)} \\ A_{21}^{(s)} & A_{22}^{(s)} \end{bmatrix} \tag{3.5.9}$$

算法的收敛速度将用 $\|A_{21}^{(s)}\|_E \to 0$ 的速度来衡量，即如能由 $\|A_{21}^{(s)}\|_E = |\varepsilon|$，导出

$$\|A_{21}^{(s+1)}\|_E = O(|\varepsilon|^p) \quad (p > 1) \tag{3.5.10}$$

对充分大的 s 成立，则称算法是 p-阶收敛的。这里 $A_{11}^{(s)}$ 是 r 阶方阵，而 $1 \leqslant r < n$。

定理 3.5.1 由实对称阵 A_1 生成的带位移的 QR 分解序列具有三阶的敛速，而位移量的选择始终用 $k_s = a_{nn}^{(s)}$。

证明 设 A_1 的依模最小的特征值 λ_n 的重数是 r，即

$$|\lambda_1| \geqslant \cdots \geqslant |\lambda_{n-r}| > |\lambda_{n-r+1}| = \cdots = |\lambda_n|, \qquad (3.5.11)$$

并且 $\lambda_{n-r+1} = \cdots = \lambda_n$，记

$$A_s = \begin{bmatrix} F_s & G_s \\ G_s^T & H_s \end{bmatrix} \}(n-r) \qquad (3.5.12)$$

据 QR方法的收敛性的一般理论可知，G_s 将以零阵为极限，记 $\lambda_i'(i=1,2,\cdots,n-r)$ 为 F_s 的特征值。$\lambda_i''(i=n-r+1, \cdots, n)$ 为 H_s 的特征值，当 s 充分大时，可设

$$A_s = \begin{bmatrix} F_s & \varepsilon k_s \\ \varepsilon k_s^T & H_s \end{bmatrix}, \quad \|k_s\|_E = 1 \qquad (3.5.13)$$

及

$$|\lambda_i' - \lambda_n| > \frac{2}{3}\left(|\lambda_{n-r} - \lambda_n|\right), \ (i=1,\cdots,n-r), \qquad (3.5.14)$$

$$\varepsilon < \frac{1}{3}|\lambda_{n-r} - \lambda_n|, \qquad (3.5.15)$$

由于 A_s 以 λ_n 为 r 重特征值，故 $A_s - \lambda_n I$ 的秩为 $(n-r)$。另一方面，$F_s - \lambda_n I$ 的 $n-r$ 个特征值为 $\lambda_i' - \lambda_n$ $(i=1,\cdots,n-r)$，它们均不为零，故 $F_s - \lambda_n I$ 的秩也是 $(n-r)$，设 $F_s - \lambda_n I$ 有三角分解

$$F_s - \lambda_n I = L_s R_s, \qquad (3.5.16)$$

作

$$L = \begin{bmatrix} L_s^{-1} & 0 \\ P_s & I \end{bmatrix} \}n-r, \qquad (3.5.17)$$

其中 P_s 满足

$$P_s(F_s - \lambda_n I) + \varepsilon k_s^T = 0. \qquad (3.5.18)$$

由于 $F_s - \lambda_n I$ 有逆，P_s 显然有解，于是有

$$L(A_s - \lambda_n I) = \begin{bmatrix} R_s & \varepsilon L_s^{-1} k_s \\ 0 & \varepsilon P_s k_s + H_s - \lambda_n I \end{bmatrix}, \qquad (3.5.19)$$

上式两端的秩均应是 $n-r$，必有

$$\varepsilon P_s k_s + H_s - \lambda_n I = 0 \qquad (3.5.20)$$

与 (3.5.18) 联立消去 P_s 得

$$-\varepsilon^2 k_s^T (F_s - \lambda_n I)^{-1} k_s + (H_s - \lambda_n I) = 0, \qquad (3.5.21)$$

$$H_s = \lambda_n I + \varepsilon^2 k_s^T (F_s - \lambda_n I)^{-1} k_s$$
$$= \lambda_n I + M_s, \qquad (3.5.22)$$

而

$$M_s = \varepsilon^2 k_s^T (F_s - \lambda_n I)^{-1} k_s, \qquad (3.5.23)$$

于是

$$\|M_s\|_E = \varepsilon^2 \|k_s^T (F_s - \lambda_n I)^{-1} k_s\|_E$$
$$\leqslant \varepsilon^2 \|k_s^T (F_s - \lambda_n I)^{-1}\|_E \|k_s\|_E$$
$$\leqslant \varepsilon^2 \|k_s^T\|_E \|(F_s - \lambda_n I)^{-1}\|_2^*$$
$$\leqslant \varepsilon^2 / \frac{2}{3} |\lambda_{n-r} - \lambda_n|$$
$$\leqslant \frac{3}{2} \cdot \frac{1}{9} |\lambda_{n-r} - \lambda_n| = \frac{1}{6} |\lambda_{n-r} - \lambda_n|. \qquad (3.5.24)$$

特别，有

$$|a_{nn}^{(s)} - \lambda_n| < \frac{3}{2} \varepsilon^2 / |\lambda_{n-r} - \lambda_n| < \frac{1}{6} |\lambda_{n-r} - \lambda_n|. \qquad (3.5.25)$$

据此，可导出

$$|\lambda_i' - a_{nn}^{(s)}| > |\lambda_i' - \lambda_n| - |\lambda_n - a_{nn}^{(s)}| > \frac{1}{2} |\lambda_{n-r} - \lambda_n| \qquad (3.5.26)$$
$$(i = 1, 2, \cdots, n-r),$$

$$|\lambda_i'' - a_{nn}^{(s)}| < |\lambda_i'' - \lambda_n| + |\lambda_n - a_{nn}^{(s)}| < 3\varepsilon^2 / |\lambda_{n-r} - \lambda_n| \qquad (3.5.27)$$
$$(i = n-r+1, \cdots, n).$$

下面我们研究对 A_s 实行一步带位移的QR算法的效果。位移量取为 $a_{nn}^{(s)}$，令

$$P_s = \begin{bmatrix} Q_s & R_s \\ S_s & T_s \end{bmatrix}, \qquad (3.5.28)$$

以使 $P_s(A_s - a_{nn}^{(s)} I)$ 成为上三角阵的正交阵，于是

* 此处用到不等式：$\|AB\|_E \leqslant \min\{\|A\|_2 \|B\|_E, \|A\|_E \|B\|_2\}$.

$$A_{s+1}=P_s(A_s-a_{nn}^{(s)}I)P_s^T$$
$$=\begin{bmatrix}F_{s+1} & G_{s+1}\\ G_{s+1}^T & H_{s+1}\end{bmatrix},$$

应有

$$S_s(F_s-a_{nn}^{(s)}I)+\varepsilon T_s k_s^T=0, \qquad (3.5.29)$$
$$G_{s+1}^T=(\varepsilon S_s k_s+T_s(H_s-a_{nn}^{(s)}I))R_s^T. \qquad (3.5.30)$$

利用$\|R_s\|_E=\|S_s\|_E$，可得

$$\|G_{s+1}^T\|_E\leqslant\varepsilon\|S_s\|_E^2+\|T_s\|_E\|S_s\|_E\|H_s-a_{nn}^{(s)}I\|_2$$

及

$$\|S_s\|=\varepsilon\|T_s k_s^T(F_s-a_{nn}^{(s)}I)^{-1}\|_E$$
$$\leqslant\varepsilon\|T_s\|_E\max_{1\leqslant i\leqslant n-r}|(\lambda_i^{-1}-a_{nn}^{(s)})^{-1}|$$
$$<2\varepsilon\|T_s\|_E/|\lambda_{n-r}-\lambda_n|. \qquad (3.3.31)$$

最后

$$\|G_{s+1}\|_E\leqslant 4\varepsilon^3\|T_s\|_E^2/|\lambda_{n-r}-\lambda_n|^2$$
$$+2\varepsilon\|T_s\|_E^2\max|\lambda_i''-a_{nn}^{(s)}|/|\lambda_{n-r}-\lambda_n|$$
$$<10\varepsilon^3\|T_s\|_E^2/|\lambda_{n-r}-\lambda_n|^2$$
$$\leqslant 10r\varepsilon^3/|\lambda_{n-r}-\lambda_n|^2.$$

这就证明了算法的三阶收敛性。由证明过程还可看出，A的依模最小的r个特征值是同时求出的，于是有重特征值的情形对于对称矩阵来说，就不再是一种缺点而是一种优点了。

5.3 对称带状阵

设A是带宽为$2m+1$的实对称阵，即

$$a_{ij}=a_{ji} \qquad i,j=1,\cdots,n,$$
$$a_{ij}=0 \qquad |i-j|>m. \qquad (3.5.32)$$

首先指出：对称带状阵在QR分解变换后仍为对称带状阵，且带宽不变。

令 $P_i(i=1,2,\cdots,n-1)$ 是初等 Hermite 阵，其中 P_i 将使 $P_i(p_{i-1}\cdots P_1A)$ 的第 i 列上三角化，且$P_i\cdots P_1A$ 与 $P_{i-1}\cdots P_1A$ 的前 $i-1$ 列是一样的，于是

$$P_{n-1}\cdots P_1A=R \tag{3.5.33}$$

是上三角阵，它的上半带宽是 $2m+1$ 而不是 $m+1$，令

$$Q=P_1\cdots P_{n-1},$$

有

$$A=QR$$

及$A_2=RQ=RP_1\cdots P_{n-1}$，依次将 $P_1,P_2\cdots,P_{n-1}$自右边连乘于 R，可见 $RP_1\cdots P_{n-1}$ 的下三角部分是带宽为 $m+1$ 的阵，由于 $RP_1\cdots P_{n-1}=QR$ 应是对称阵，故 QR 实际是带宽为 $2m+1$ 的对称阵。

在讨论对称带状阵的带位移的 QR 分解时，位移量 k_s 的选择，仍可按 §4 中所述的一般原则进行。本段所要讨论的是节约内存的最大可能性。

由于 A_1 是对称带状阵，可以只输入其上三角部分。为输入方便，可以把A的非零元 $a_{i,k+i}$ 放在二维数组$B[1:n;0:m]$的$B[i,k]$单元中，这里 $i=1,2,\cdots,n,k=0,1,\cdots,\min(n-i,m)$。由 A_1 计算 A_2 时，主要经过以下两个步骤：

(i) 由 A_1 计算 $R=P_{n-1}\cdots P_1A_1$；

(ii) 由R计算$A_2=RP_1\cdots P_{n-1}$。

由于 $P_iP_{i-1}\cdots P_1A_1$ 不改变 $P_{i-1}\cdots P_1A$ 的前 $i-1$ 行，所以$P_i\cdots P_1A_1$ 的第 i 行的输出就是 $P=P_{n-1}\cdots P_1A_1$ 的第 i 行的输出，由于 $P_i\cdots P_1A$ 的第 i 行的宽度是 $2m+1$ 而不是 $m+1$，这样，用它去冲掉 A_1 的第 i 行是有困难的。但是如果注意到在由R 计算A_2 时，有些不必记录的中间结果可以及时剔除掉，以便尽可能的节省存贮量和计算量。

以$m=2$，$n=7$为例，我们将通过比较R和RP_1的存贮情况来描述这个问题的计算与存贮的策略

$$
\begin{array}{cc}
R & RP_1 \\
\left(\begin{array}{ccccccc}
\times & \times & \times & \times & \times & \bigcirc & \bigcirc \\
 & \times & \times & \times & \times & \times & \bigcirc \\
 & & \times & \times & \times & \times & \times \\
 & & & \times & \times & \times & \times \\
 & & & & \times & \times & \times \\
 & & & & & \times & \times \\
 & & & & & & \times
\end{array}\right)
&
\left(\begin{array}{ccccccc}
\triangle & \bullet & \bullet & \times & \times & & \\
\triangle & \otimes & \otimes & \times & \times & \times & \\
\triangle & \otimes & \otimes & \times & \times & \times & \times \\
 & & & \times & \times & \times & \times \\
 & & & & \times & \times & \times \\
 & & & & & \times & \times \\
 & & & & & & \times
\end{array}\right)
\end{array}
$$

R在右乘以P_1后，它只影响R的前三列，即画有△、●和⊗的那些元素。注意，画有△的元素，在以后继续右乘P_2，…时将不再变化，即它们是A_2的第1列的最后输出信息；画有⊗的元素在RP_1继续右乘P_2…时，它们将发生变化并且对于形成A_2的第2、3列在对角线下的元素是有贡献的；画有●的元素，在以后的计算中，虽然也要发生变化，但它们对形成A_2的对角线以下的第2列以后的元素却没有贡献。仿此分析可知，在R中真正对形成A_2的各列在对角线以下的元素有贡献的只是R中位于第$1,2,\cdots,m+1$条对角线上的那些元素。可见R在右乘以P_1之后，RP_1的第1行中的前$m+1$个元素可以用RP_1的第1列中画有△的元素冲掉。同样，当右乘P_2以后，可用RP_1P_2的第2列在对角线以下的元素冲掉RP_1P_2的第2行对角线以右的那些元素。注意，当

$$P_{m+1}\cdots P_1A_1$$

算出以后，前$m+1$列便已经上三角化，而后继的流动矩阵

$$P_k\cdots P_{m+2}(P_{m+1}\cdots P_1A_1)$$

的前$m+1$行和$m+1$列与$P_{m+1}\cdots P_1A_1$的前$m+1$行列是不变

化的，因此

$$P_{m+1}\cdots P_1 A_1 P_1$$

的第1列就是最后输出的第1列。在这之后，P_1 的信息就不再有用而可以退出存区了。因此，计算如果按

$$\cdots(P_{m+2}(P_{m+1}\cdots P_1 A_1)P_1)P_2\cdots \qquad (3.5.34)$$

的次序进行，就可以不必把 $P_1, \cdots P_{n-1}$ 的全部信息存贮起来，而只需准备一个存贮

$$P_1, P_2, \cdots, P_{m+1}$$

的存区，记为 $\overline{W}$ 存区，其容量为 $(m+1)^2$。此外，还应有对

$$P_{m+i}\cdots P_{m+2}(P_{m+1}\cdots P_1 A)P_1\cdots P_i$$

进行加工的辅助存区。其一，实现由 A_1 计算 R，称为辅助存区Ⅰ；其二，实现由 R 计算 A_2 称为辅助存区Ⅱ，二者的容量均为 $(m+1)\times(2m+1)$。

由 A_1 计算 A_2 的主要步骤如下：

算法 3.5.1

1）辅区Ⅰ清零，辅区Ⅱ清零；把对称带状阵 A 的上三角部分元素 $(i, i+k)$ 依压缩形式放入二维数组 $B[1:n, 0:m]$ 的 $B[i,k]$ 中，$(i=1,2,\cdots n, k=0,1,\cdots,\min(n-i,m))$，以 $n=6$，$m=2$ 为例，则 A、B 的初态各为

$$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} & & & \\ a_{12} & a_{22} & a_{23} & a_{24} & & \\ a_{13} & a_{23} & a_{33} & a_{34} & a_{35} & \\ & a_{24} & a_{34} & a_{44} & a_{45} & a_{46} \\ & & a_{35} & a_{45} & a_{55} & a_{56} \\ & & & a_{46} & a_{56} & a_{66} \end{pmatrix},$$

$$B=\begin{pmatrix} b_{10} & b_{11} & b_{12} \\ b_{20} & b_{21} & b_{22} \\ b_{30} & b_{31} & b_{32} \\ b_{40} & b_{41} & b_{42} \\ b_{50} & b_{51} & 0 \\ b_{60} & 0 & 0 \end{pmatrix}=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{22} & a_{23} & a_{24} \\ a_{33} & a_{34} & a_{35} \\ a_{44} & a_{45} & a_{46} \\ a_{55} & a_{56} & 0 \\ a_{66} & 0 & 0 \end{pmatrix};$$

2）确定位移量 k_s。欲求A的依模最小的特征值时，可取$k_s=0$；欲求A的在某 t 值附近的特征值时，可取$k_s=t$。并置$b_{i0}\leftarrow(b_{i0}-k_s)\ (i=1,\cdots,n)$；

3）为辅区 I 准备初态。即把B的前$m+1$行抄入辅区 I 的对角线右侧的各行，并使下三角部分与上三角部分对称。以$m=2$为例，此时辅区 I 的初态为

辅 助 存 区 I（初态）

$$\begin{pmatrix} b_{10}-k_s & b_{11} & b_{12} & 0 & 0 \\ b_{11} & b_{20}-k_s & b_{21} & b_{22} & 0 \\ b_{12} & b_{21} & b_{30}-k_s & b_{31} & b_{32} \end{pmatrix}.$$

以下按（3.5.34）的计算次序实现算法。

4）对 $i=1,2,\cdots,m+n$ 做下列循环：

当$i<n$时做 4.1)，否则转 4.4)；

4.1）用辅区 I 的第 1 列做单位向量 $\mathbf{w}_i$ 及矩阵 $P=I-2\mathbf{w}_i\mathbf{w}_i^T$，用$P$左乘辅区 I；

4.2）当 $i\leqslant m+1$ 时做 4.3)，否则转 4.4)；

4.3）送 $\mathbf{w}_i$ 到W存区的第 i 行；辅区 I 的第 1 行送入辅区 II 的第 i 行对角线右侧；转 4)；

4.4）用W的第 1 行 $\mathbf{w}$做 $P=I-2\mathbf{w}\mathbf{w}^T$，用$P$右乘辅区 II，送辅区 II 第 1 列到$B$的 $i-m-1$ 行，辅区 II 上移 1 行，左移 1 列，末行清零，辅区 I 的第 1 行前 $m+1$ 个元素送入辅区 II 末

行对角线右侧，辅区 I 上移 1 行，左移 1 列，末行清零；W区上移一行；

4.5）当 $i<n$ 时，做 4.6），否则转 4）

4.6）$\mathbf{w}_i$ 送入W区末行；B 的第 i 行送辅区 I 末行对角线右侧，B 的第 i 条对角线（走向向上），即 b_{i1}，$b_{i-1,2}$，$\cdots$ $b_{i-m+1,m}$ 分别送入辅区 I 中 $(m+1,m)$，$(m+1,m-1)$，$\cdots$，$(m+1,1)$ 的位置；转 4）。

以上步骤完成时，在数组 B 中得到的将是 A 的带位移的 QR 分解的压缩存贮形式。

第三章习题

3.1 将矩阵

$$\begin{pmatrix} 0 & 1 & & \\ 1 & 0 & 0.01 & \\ & 0.01 & 0 & 1 \\ & & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0.01 & & \\ 0.01 & 1 & 0.01 & \\ & 0.01 & 1 & 0.01 \\ & & 0.01 & 1 \end{pmatrix}$$

各做两次带位移的 QR 分解，位移量取为矩阵右下角的元素。

3.2 试证，矩阵

$$P = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

的 QR 分解序列是不变的。

3.3 求矩阵

$$\Lambda = \begin{pmatrix} & & \lambda_1 \\ & \lambda_2 & \\ \lambda_3 & & \end{pmatrix}$$

的 QR 分解序列，它收敛否？

3.4 不可约的 Hessenberg 阵是非减次的，证明之。

3.5　构造矩阵 A 使其 LR 分解是发散的。

3.6　证明三对角对称阵

$$S=\begin{pmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & e_3 & & \\ & e_3 & \ddots & & \\ & & \ddots & \ddots & e_n \\ & & & e_n & d_n \end{pmatrix}$$

的所有元素满足，$|d_i|, |e_i| \leqslant \max_i |\lambda_i|$，这里 λ_i 是矩阵 S 的第 i 个特征值。又 S 的QR分解中的上三角阵R的元素满足 $|r_{ij}| \leqslant 2\max |\lambda_k|$。

3.7　设 B 是实对称矩阵，$B=QR$ 是其 QR 分解。

试证

$$|r_{nn}| \geqslant \min |\lambda_i(B)|。$$

3.8　设A是不可约的奇异的 Hessenberg 阵，A_2 是A的QR分解的第2项。试证 A_2 的最后一行是零行。设 A 是不可约的且以$\lambda=0$为 m 重特征值的 Hessenberg 阵，试推广所证的结果。

3.9　设 $\det(A)\neq 0$，则存在排列阵 P 使 PA 有 LR 分解，且其 LR 分解中的单位下三角阵的元素依模不超过 1。

3.10　设 $A_s \to I$，$A_s=Q_sR_s$ 是 A_s 的QR分解。证明，$Q_s \to I$，$R_s \to I$。

3.11　依公式（3.4.3）～(3.4.9）编制一个求上 Hessenberg 阵的带两步位移的 QR分解程序，并用它计算

$$A_1=\begin{pmatrix} 5 & -2 & -5 & -1 \\ 1 & 0 & -3 & 2 \\ 0 & 2 & 2 & -3 \\ 0 & 0 & 1 & -2 \end{pmatrix}$$

的A_3。

3.12　编制一个求带行交换的计算上 Hessenberg 阵的行列式的程序，并用它计算

$$\begin{pmatrix} 1 & 0.25 & & \\ 2 & 1 & 0.25 & \\ & 2 & 1 & 0.25 \\ & & 2 & 1 \end{pmatrix}$$

的行列式。

3.13　按本章第5节所述的方法编制一个求对称带状阵的全部特征值的程序，并用它计算

$$\begin{pmatrix} 4 & 2 & 1 & & \\ 2 & 4 & 2 & 1 & \\ 1 & 2 & 4 & 2 & 1 \\ & 1 & 2 & 4 & 2 \\ & & 1 & 2 & 4 \end{pmatrix}$$

的全部特征值。

3.14　设A是正定三对角对称阵，A_s是A的 QR 分解序列的第s项，设记A_s的对角元素为：

$$\alpha_{s1},\ \alpha_{s2},\cdots,\ \alpha_{sn}$$

而次对角元记为：

$$\beta_{s1},\quad \beta_{s2},\cdots,\beta_{sn-1}.$$

试证，$\alpha_{s+1,1},\ \cdots,\ \alpha_{s+n,n}$；$\beta^2_{s+1,1},\ \cdots,\ \beta^2_{s+1,n-1}$可由$\alpha_{s1},\cdots,\alpha_{sn}$，$\beta^2_{s1}\cdots,\ \beta^2_{sn}$经有限次有理运算得到。

第四章　广义特征值问题

§1　引　　言

设 A 和 B 均为 n 阶矩阵，广义特征值问题就是求满足矩阵方程

$$Ax = \lambda Bx \tag{4.1.1}$$

的数 λ 及非零向量 x，它们分别称为广义特征值问题 $A-\lambda B$ 的特征值和特征向量。当然应该把 $N(A)\cap N(B)$ 中的非零向量（如果存在的话）排除在外。因为对于这些向量，任意数 λ 都是特征值，显然这是毫无意义的。排除了这种情形，特征值 λ 为特征多项式

$$p_n(\lambda) = \det(A-\lambda B) \tag{4.1.2}$$

的零点。

现在，我们在前章的基础上，对广义特征值问题的计算方法作初步分析：考虑一下解决这一问题有些什么困难与矛盾；设想一下解决这一问题有哪些途径。

最为自然的想法是：如果 B 为非异，那末方程（4.1.1）就等价于方程

$$B^{-1}Ax = \lambda x. \tag{4.1.3}$$

这样问题（4.1.1）就化为 $B^{-1}A$ 的标准特征值问题。只要 B 关于求逆运算并非病态，那末就可利用任何一种标准特征值问题的计算方法来求解问题（4.1.3）。

在工程计算中，常常会碰到 A 和 B 均为对称正定的情形（或仅其中之一为对称正定）。譬如，在结构系统的振动问题中，A

为刚度矩阵，B 为质量矩阵，$\lambda=\omega^2$（ω 为自振频率）就属于这种情形。这时问题（4.1.1）就称为对称广义特征值问题。由于 B 是对称正定阵，所以对它可以作 Cholesky分解：

$$B=LL^T$$

这里 L 是对角元皆为正的下三角阵。众所周知，这一分解是十分稳定的。于是（4.1.1）就等价于

$$L^{-1}AL^{-T}(L^T\boldsymbol{x})=\lambda(L^T\boldsymbol{x}).$$

若令

$$\boldsymbol{y}=L^T\boldsymbol{x}, \qquad (4.1.4)$$

$$C=L^{-1}AL^{-T}, \qquad (4.1.5)$$

则（4.1.1）问题即等价于标准对称特征值问题：

$$C\boldsymbol{y}=\lambda\boldsymbol{y}. \qquad (4.1.6)$$

这样前面讲的关于对称特征值问题的计算方法都可用来求解（4.1.6）。例如，可用TRED2将C化为对称三对角形，然后利用TQL2求出C的全部特征值和特征向量，最后利用 REBAK求出原问题的特征向量。

然而，在以上讨论的各种计算方案中，蕴含着许多矛盾。首先，如果B 对于求逆运算为病态，那末此时就不能较为精确地计算 $B^{-1}A$。从而必然影响特征值、特征向量的精确计算。事实上，当B 关于求逆运算为病态时，一个本质的困难在于：矩阵 B 的最小奇异值σ_n相对而言较小，因而问题（4.1.3）必有一些相对而言较大的特征值，而它们对于 B的扰动是十分敏感的。例如。设

$$A=\begin{pmatrix}1 & 2\\ 2 & 0\end{pmatrix}, \quad B=\begin{pmatrix}1+\varepsilon & 1\\ 1 & 1\end{pmatrix},$$

其中 $\varepsilon>0$。则

$$\det(A-\lambda B)=\varepsilon\lambda^2+3\lambda-4,$$

故 $A-\lambda B$ 的特征值是

$$\lambda_{1,2}=\frac{-3\pm\sqrt{9+16\varepsilon}}{2\varepsilon}.$$

当$\varepsilon\to 0$时，$A-\lambda B$ 的特征值趋于 $\frac{4}{3}$ 和 ∞。这说明，ε 很小的变化，即对B 很小的扰动，就能引起第二个特征值很大的改变。

为了解决这一矛盾，显然应避免求逆运算，QZ 算法就是基于这一想法产生的一种算法。

前面，已介绍了求解标准特征值问题的 QR 算法，其基本思想是：采用一系列（无穷个）酉相似变换将原始问题化为上三角形或拟上三角形。由此我们得到启发，是否能将这种想法移植过来，以求解广义特征值问题。

因为对A和 B 进行相同的酉等价变换并不改变原始问题的特征值。所以人们想到利用酉等价变换同时将A、B化为上三角形$\tilde{A}$和$\tilde{B}$。这样，问题(4.1.1)的特征值即可由$\tilde{A}$、$\tilde{B}$对角元之比给出，而整个计算过程并不涉及求逆运算。但这一算法也有缺点，它在计算过程中要破坏原始矩阵的性质，如对称性、稀疏性等等。

对于对称的广义特征值问题，固然可以化为对称的标准特征值问题。可是，实际工程问题中的矩阵A和 B 往往是对称带形矩阵，带宽常常很小。而将（4.1.1）化为（4.1.6）意味着带形结构的破坏。因为一般而言矩阵C 将为一满阵。不充分利用A、B稀疏带状的特点，而间接求解（4.1.6），这样的方案显然是低效的。针对A、B为稀疏带状的性质，**Peters** 和**Wilkinson** 提出了直接求解对称广义特征值问题的方法。其基本思想与标准特征值问题的 **Givens-Householder** 方法是类同的。简言之，即利用 **Sturm** 序列：

$$P_0(\lambda)\equiv 1,$$

$$P_r(\lambda)\equiv\det(A^{[r]}-\lambda B^{[r]}) \qquad (4.1.7)$$

的同符号数，结合二分法求得满足给定转度的特征值。而确定sturm序列（4.1.7）的同符号数时，可以充分利用A、B的稀疏带状性质。（4.1.7）式中$A^{[r]}$表示矩阵A的r阶首主子矩阵。

在工程计算中，对于振动问题，往往只需求$A-\lambda B$最小的几个特征值及其相应的特征向量（相当于求系统的几个最低的固有频率及其振型）。这时对（4.1.6）采用子空间迭代法也是有利的。采用子空间迭代法求解（4.1.6）时，$C=L^{-1}AL^{-T}$不必真正形成，而只要存贮其因子L和矩阵A。因为在子空间迭代法中，只用到矩阵A与向量的乘积，以及求解系数矩阵为L和L^T的方程组。其次，它又可以利用矩阵A的稀疏性，以提高计算效率。

§2 QZ 算 法

按前节的设想，QZ 算法可按以下三个步骤进行：

（1） 首先利用正交三角化算法（例如可采用 **Householder** 方法或修正的 **Gram-Schmidt** 算法）将B化为上三角形$\tilde{B}$，同时将这些变换施行于A化为$\tilde{A}$。如$n=6$时有

$$\begin{array}{cccccc} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \end{array} \qquad \begin{array}{cccccc} \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{array};$$

$$\tilde{A} \qquad\qquad\qquad\qquad \tilde{B}$$

（2）利用平面旋转阵逐步把矩阵$\tilde{A}$化为上 **Hessenberg** 阵，但要求这些平面旋转阵同时作用$\tilde{B}$后仍保持为上三角阵。消去$\tilde{A}$

次对角线下的元素，其顺序为：$(n,1)$，$(n-1,1)$，…，$(3,1)$；$(n,2)$，$(n-1,2)$，…，$(4,2)$；$(n,n-2)$一元，消去 $\bar{A}$的(i,j)一元，可通过左乘平面旋转阵

$$Q_{i-1,i}=\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \cos\varphi_i & \sin\varphi_i & & \\ & & & -\sin\varphi_i & \cos\varphi_i & & \\ & & & & & 1 & \\ & & & & & & \ddots \\ 0 & & & & & & 1 \end{pmatrix}\begin{matrix} \\ \\ \\ i-1 \\ i \\ \\ \\ \\ \end{matrix} \qquad (4.2.1)$$

$$\qquad\qquad i-1 \qquad i$$

来实现。而$Q_{i-1,i}$作用于$\tilde{B}$时，在$\tilde{B}$的$(i,i-1)$位置上(即次对角线上)引入了一个非零元。但这个非零元可以通过右乘另一平面旋转阵

$$Z_{i-1,i}=\begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \cos\theta_i & \sin\theta_i & & \\ & & & -\sin\theta_i & \cos\theta_i & & \\ & & & & & 1 & \\ & & & & & & \ddots \\ 0 & & & & & & 1 \end{pmatrix}\begin{matrix} \\ \\ \\ i-1 \\ i \\ \\ \\ \\ \end{matrix} \qquad (4.2.2)$$

$$\qquad\qquad i-1 \qquad i$$

消去。然而这个矩阵右乘 $Q_{i-1,i}\bar{A}$ 并不破坏已经得到的零元素以上例来说明，第一步变换过程如下：

$$\begin{matrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \end{matrix} \qquad \begin{matrix} \times & \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{matrix}$$

$$Q_{56}\bar{A}Z_{56} \qquad\qquad\qquad Q_{56}\tilde{B}$$

$$\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & 0 & \times
\end{array}$$

$$Q_{56}\tilde{B}Z_{56}$$

第二步变换过程如下：

$$\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times
\end{array}
\qquad
\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & 0 & \times
\end{array}$$

$$Q_{45}Q_{56}\tilde{A}Z_{56} \qquad\qquad Q_{45}Q_{56}\tilde{B}Z_{56}$$

$$\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times
\end{array}
\qquad
\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & 0 & \times
\end{array}$$

$$Q_{45}Q_{56}\tilde{A}Z_{56}Z_{45} \qquad\qquad Q_{45}Q_{56}\tilde{B}Z_{56}Z_{45}$$

如此继续执行以上步骤，最后 $\tilde{A}$ 化为上 Hessenberg 阵而 $\tilde{B}$ 仍保持上三角形。我们把这两个矩阵仍分别记作 $\tilde{A}$，$\tilde{B}$。

（3）因为

$$\tilde{A} = UAV,$$

$$\tilde{B} = UBV,$$

这里U，$V\in R^{n\times n}$为第一类正交阵．因为

$$\det(\tilde{A}-\lambda\tilde{B})=\det(U)\cdot\det(V)\cdot\det(A-\lambda B)$$

$$=\det(A-\lambda B)。$$

由此可见，问题（4.1.1）应与问题

$$\tilde{A}x=\lambda\tilde{B}x \tag{4.2.3}$$

有相同的特征值。另外，若x为问题（4.2.3）的特征向量，则Vx为问题（4.1.1）对应的特征向量。因此，现在只要讨论广义特征值问题（4.2.3）的计算方法就行了。

现设$\tilde{B}$为非异，而$\tilde{C}\equiv\tilde{A}\tilde{B}^{-1}$。由于$\tilde{A}$为上 Hessenberg 阵而$\tilde{B}$为上三角形，故$\tilde{C}$也为上 Hessenberg 阵。根据§1的分析，现在的目标是要把$\tilde{A}$与$\tilde{B}$同时化约为上三角阵，也就等价于将$\tilde{C}$化约为上三角阵。

鉴于以上分析，我们想到利用带原点位移的 QR 算法。若将它施于$\tilde{C}$，得到矩阵序列$\{C_k\}$，则在一定条件下，该矩阵序列应趋于一上三角阵。具体地说，即

$$\begin{cases}\tilde{C}\equiv C_1,\\ Q_k(C_k-\kappa_k I)\equiv R_k, \quad (k=1,2,\cdots)\\ R_kQ_k^H+\kappa_k I\equiv C_{k+1},\end{cases} \tag{4.2.4}$$

且有

$$C_{k+1}=Q_kC_kQ_k^H, \tag{4.2.5}$$

然而我们的目的并不是要从C_k直接产生Q_k，再进而求出C_{k+1}。假定A_k与B_k满足$A_kB_k^{-1}\equiv C_k$，且A_k与B_k仍分别为上 Hessenberg 阵与上三角阵，那末可以设想这样的计算方案：避开直接计算B_k^{-1}，间接地求出Q_k，继而构造A_{k+1}和B_{k+1}：

$$A_{k+1}=Q_kA_kV_k, \tag{4.2.6}$$

$$B_{k+1}=Q_kB_kV_k。$$

这里V_k为酉阵，它要保证A_{k+1}为上 Hessenberg 阵而B_{k+1}为上三角阵。于是

$$
\begin{aligned}
C_{k+1} &\equiv A_{k+1} B_{k+1}^{-1} \\
&= (Q_k A_k V_k)(Q_k B_k V_k)^{-1} \\
&= Q_k C_k Q_k^H.
\end{aligned}
\tag{4.2.7}
$$

如何避开直接计算B_k^{-1}而间接地求出Q_k呢？由定理3.3.1知：对（4.2.7）而言，如果C_{k+1}为不可约的上 Hessenberg 阵，则C_{k+1}与Q_k唯一地（除模为1的常数因子外）由Q的第一行决定。这样我们就很自然地想到隐位移 QR 算法，为避免复数运算又可采用二步QR算法。

据前章分析知，为实现二步 QR 算法，首先应找到一酉阵P_0，它的第一行应与矩阵

$$
\varphi_2(C_k) \equiv (C_k - \chi_k I)(C_k - \chi_{k+1} I) \tag{4.2.8}
$$

的第一列成比例，其中$\chi_{k+1} = \bar{\chi}_k$。而

$$
\varphi_2(C_k)\boldsymbol{e}_1 = \begin{pmatrix} (\gamma_{11}^{(k)})^2 + \gamma_{21}^{(k)}\gamma_{12}^{(k)} \\ \gamma_{21}^{(k)}\gamma_{11}^{(k)} + \gamma_{22}^{(k)}\gamma_{21}^{(k)} \\ \gamma_{32}^{(k)}\gamma_{21}^{(k)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} - (\chi_k + \chi_{k+1}) \begin{pmatrix} \gamma_{11}^{(k)} \\ \gamma_{21}^{(k)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}
$$

$$
+ \chi_k \chi_{k+1} \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.2.9}
$$

由此可见，$\varphi_2(C_k)\boldsymbol{e}_1$仅用到$C_k$的前二列。但因$B_k$是上三角形，故$B_k^{-1}$亦为上三角形。于是$C_k$的前二列为

$$(a_1^{(k)}, a_2^{(k)})\begin{pmatrix} \beta_{11}^{(k)} & \beta_{12}^{(k)} \\ 0 & \beta_{22}^{(k)} \end{pmatrix}^{-1}. \tag{4.2.10}$$

这样，为了计算P_0，只需计算B_k的二阶首主子矩阵的逆即可。显然，这要比计算B_k^{-1}的效果好得多。又由前章知，因为$\varphi_2(C_k)$的第一列至多前三个元素不为零，故P_0不妨取作如下形状：

$$P_0 = \begin{pmatrix} R_0 & 0 \\ 0 & I_{n-3} \end{pmatrix}, \tag{4.2.11}$$

其中R_0是一个三阶的初等镜像变换阵。设$\varphi_2(C_k)e_1 \equiv \tilde{q}_1$，则$P_0^T$的第一列即为$\hat{q}_1 = \pm\dfrac{\tilde{q}_1}{\|\tilde{q}_1\|}$，即$P_0$应具有这样的性质：$P_0^T e_1 = \hat{q}_1$或$\tilde{q}_0\tilde{q}_1 = \sigma e_1 (\sigma = \pm\|\tilde{q}_1\|)$。当然这样的$P_0$是能构造出来的。$P_0A_k$和$P_0B_k$应具有如下形状（仍以$n=6$为例）

$$\begin{matrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{matrix}; \qquad \begin{matrix} \times & \times & \times & \times & \times & \times \\ \otimes & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{matrix}.$$

$$P_0A_k \qquad\qquad\qquad\qquad P_0B_k$$

根据以上设想方案，为了求出A_{k+2}和B_{k+2}，需找到这样的正交阵U_k和V_k，使$U_kP_0A_kV_k$为上Hessenberg阵，$U_kP_0B_kV_k$为上三角阵。同时，还要保证$U_kP_0 = Q_k'$与P_0有相同的第一行，这样Q_k'与Q_k即有相同的第一行。据定理3.3.1知，$Q_k' \equiv Q_k \cdot U_k$和$V_k$的形成步骤如下：

（1）首先将P_0B_k按下面方式化为上三角阵（以$n=6$为例）

选取初等镜像变换阵Z_0，使$P_0B_kZ_0$中（3,1）元和(3，2)

元变为零．然后再用平面旋转阵 $R_{12}=Z_0'$ 使 $P_0B_kZ_0Z_0'$ 中的（2，1）元变为零．于是 $P_0A_kZ_0Z_0'$ 和 $P_0B_kZ_0Z_0'$ 就有如下的形状：

$$
\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\otimes & \times & \times & \times & \times & \times \\
\otimes & \times & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times
\end{array}
\qquad
\begin{array}{cccccc}
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & 0 & \times
\end{array}
$$

$$P_0A_kZ_0Z_0' \qquad\qquad P_0B_kZ_0Z_0'$$

请注意这里Z_0和Z_0'的形状为：

$$
\begin{array}{cccccc}
\times & \times & \times & 0 & 0 & 0 \\
\times & \times & \times & 0 & 0 & 0 \\
\times & \times & \times & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}
\qquad
\begin{array}{cccccc}
\times & \times & 0 & 0 & 0 & 0 \\
\times & \times & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

$$Z_0 \qquad\qquad Z_0'$$

（2）选取一初等镜象变换阵P_1，使$P_0A_kZ_0Z_0'$中打圈位置上的元素变为零，P_1的形状如下：

$$
\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & \times & \times & \times & 0 & 0 \\
0 & \times & \times & \times & 0 & 0 \\
0 & \times & \times & \times & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

$$P_1$$

于是$P_1P_0A_kZ_0Z_0'$和$P_1P_0B_kZ_0Z_0'$有如下形状：

$$
\left[\begin{array}{c|ccccc}
\times & \times & \times & \times & \times & \times \\
\hline
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times
\end{array}\right]
\qquad
\left[\begin{array}{c|ccccc}
\times & \times & \times & \times & \times & \times \\
\hline
0 & \times & \times & \times & \times & \times \\
0 & \otimes & \times & \times & \times & \times \\
0 & \otimes & \otimes & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & 0 & \times
\end{array}\right]
$$

(3)对划去第一行、第一列后的子矩阵重复(1)(2)步骤。

(4)当算法进行到第$n-2(=4)$次循环第(1)步时，这二个矩阵呈如下形状：

$$
\left[\begin{array}{ccc|ccc}
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
\hline
0 & 0 & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & \otimes & \times & \times
\end{array}\right]
\qquad
\left[\begin{array}{ccc|ccc}
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & \times & \times & \times & \times \\
\hline
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & 0 & \times
\end{array}\right],
$$

选取一平面旋转阵P_4，使左边矩阵中打圈位置上的元素变为零。P_4的形状如下：

$$
\begin{array}{cccccc}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & \times & \times
\end{array}.
$$

P_4 左乘这二个矩阵后有如下形状：

$$
\left(\begin{array}{ccc|ccc}
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
\hline
0 & 0 & \times & \times & \times & \times \\
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times
\end{array}\right)
\qquad
\left(\begin{array}{ccc|ccc}
\times & \times & \times & \times & \times & \times \\
0 & \times & \times & \times & \times & \times \\
0 & 0 & \times & \times & \times & \times \\
\hline
0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times \\
0 & 0 & 0 & 0 & \otimes & \times
\end{array}\right)
$$

（5）选取一平面旋转阵Z_4，右乘右边那个矩阵，使其中打圈位置上的元素变为零。同时Z_4右乘左边的矩阵，显然它仍保持上 **Hessenberg** 阵的形式。至此化约完成，即A_{k+2}和B_{k+2}均已求得。

以上全部过程构成了所谓二步 QZ 算法。实际上，它是二步 QR 算法的一种推广。

现在进一步讨论计算细节。根据以上分析。我们知道，对（4.2.3）施行二步QZ 算法，首先应找到正交阵P_0，它的第一行应与矩阵$\varphi_2(C_k)$的第一列成比例。为此应找出C_k的前二列。据（4.2.10）知，C_k的前二列应为：

$$
\begin{pmatrix}
\alpha_{11}^{(k)}/\beta_{11}^{(k)} & \alpha_{12}^{(k)}/\beta_{22}^{(k)}-\alpha_{11}^{(k)}\beta_{12}^{(k)}/\beta_{11}^{(k)}\beta_{22}^{(k)} \\
\alpha_{21}^{(k)}/\beta_{11}^{(k)} & \alpha_{22}^{(k)}/\beta_{22}^{(k)}-\alpha_{21}^{(k)}\beta_{12}^{(k)}/\beta_{11}^{(k)}\beta_{22}^{(k)} \\
 & \alpha_{32}^{(k)}/\beta_{22}^{(k)} \\
0 & 0 \\
\vdots & \vdots \\
0 & 0
\end{pmatrix}. \qquad (4.2.12)
$$

为了确定位移量κ_k和κ_{k+1}，则应确定矩阵C_k的尾二阶主子矩阵，我们把它近似地取作：

$$
\begin{pmatrix}
\alpha_{n-1,n-1}^{(k)} & \alpha_{n-1,n}^{(k)} \\
\alpha_{n,n-1}^{(k)} & \alpha_{nn}^{(k)}
\end{pmatrix}
\begin{pmatrix}
\beta_{n-1,n-1}^{(k)} & \beta_{n-1,n}^{(k)} \\
0 & \beta_{nn}^{(k)}
\end{pmatrix}^{-1}. \qquad (4.2.13)
$$

位移量κ_k和κ_{k+1}为矩阵（4.2.13）的二个特征值。据二次方程根与系数的关系即得：

$$\begin{cases} \kappa_k+\kappa_{k+1}=\alpha_{nn}^{(k)}/\beta_{nn}^{(k)}+\alpha_{n-1,n-1}^{(k)}/\beta_{n-1,n-1}^{(k)} \\ \qquad -\alpha_{n,n-1}^{(k)}\beta_{n-1,n}^{(k)}/(\beta_{n-1,n-1}^{(k)}\cdot\beta_{nn}^{(k)}), \\ \kappa_k\cdot\kappa_{k+1}=(\alpha_{n-1,n-1}^{(k)}\alpha_{nn}^{(k)}-\alpha_{n,n-1}^{(k)}\alpha_{n-1,n}^{(k)})/(\beta_{n-1,n-1}^{(k)}\beta_{nn}^{(k)}). \end{cases} \tag{4.2.14}$$

将（4.2.12）和（4.2.14）代入（4.2.9）即得$\tilde{q}_1$前三个非零分量：

$$\begin{cases} \alpha_{10}=\left[\left(\dfrac{\alpha_{n-1,n-1}^{(k)}}{\beta_{n-1,n-1}^{(k)}}-\dfrac{\alpha_{11}^{(k)}}{\beta_{11}^{(k)}}\right)\left(\dfrac{\alpha_{nn}^{(k)}}{\beta_{nn}^{(k)}}-\dfrac{\alpha_{11}^{(k)}}{\beta_{11}^{(k)}}\right)-\dfrac{\alpha_{n,n-1}^{(k)}\alpha_{n-1,n}^{(k)}}{\beta_{nn}^{(k)}\beta_{n-1,n-1}^{(k)}}\right. \\ \qquad \left.+\dfrac{\alpha_{n,n-1}^{(k)}\beta_{n-1,n}^{(k)}}{\beta_{nn}^{(k)}\beta_{n-1,n-1}^{(k)}}\cdot\dfrac{\alpha_{11}^{(k)}}{\beta_{11}^{(k)}}\right]\dfrac{\beta_{11}^{(k)}}{\alpha_{21}^{(k)}}+\dfrac{\alpha_{12}^{(k)}}{\beta_{22}^{(k)}}-\dfrac{\alpha_{11}^{(k)}\beta_{12}^{(k)}}{\beta_{11}^{(k)}\beta_{22}^{(k)}}, \\ \alpha_{20}=\dfrac{\alpha_{22}^{(k)}}{\beta_{22}^{(k)}}+\dfrac{\alpha_{11}^{(k)}}{\beta_{11}^{(k)}}-\dfrac{\alpha_{21}^{(k)}\beta_{12}^{(k)}}{\beta_{11}^{(k)}\beta_{22}^{(k)}}-\dfrac{\alpha_{nn}^{(k)}}{\beta_{nn}^{(k)}}-\dfrac{\alpha_{n-1,n-1}^{(k)}}{\beta_{n-1,n-1}^{(k)}} \\ \qquad +\dfrac{\alpha_{n,n-1}^{(k)}\beta_{n-1,n}^{(k)}}{\beta_{nn}^{(k)}\beta_{n-1,n-1}^{(k)}}, \\ \alpha_{30}=\dfrac{\alpha_{32}^{(k)}}{\beta_{22}^{(k)}}. \end{cases} \tag{4.2.15}$$

注意，首先，$\tilde{q}_1$的每个分量我们都乘以$\beta_{11}^{(k)}/\alpha_{21}^{(k)}$，这对确定$P_0$并无影响。确定$P_0$的具体过程，见前章§3；其次，为了程序设计上的原因，把$\tilde{q}_1$的三个非零分量分别记作α_{10}，α_{20}，α_{30}，把P_0A_k第一列的元素记作α_{i1}，把$P_1P_0A_kZ_0Z_0'Z_1Z_1'$的第二列记作α_{i2}等等。至此，我们即可构成二步QZ算法。

算法 4.2.1 设给定矩阵$A, B\in R^{n\times n}$而B为非异矩阵，用QZ算法求（4.1.1）的全部特征值。用此算法求解的步骤如下：

1）利用正交三角化算法将B化为上三角阵$\tilde{B}$，同时将A化为$\tilde{A}$。置$A=\tilde{A}$，$B=\tilde{B}$；

2）利用平面旋转阵将 A 化为上 Hessenberg 阵，同时保持 B 为上三角形；

3）对 $j=n,\ n-1,\cdots,1$.

3.1）对 $k=1,2,\cdots,m$（允许的最大迭代次数）.

3.1.1）按（4.2.15）计算 α_{10}，α_{20}，α_{30}.

3.1.2）对 $i=1,2,\cdots,n-2$；

3.1.2.1）确定 P_{i-1} 消去 $\alpha_{i+1,i-1}^{(k)}$，$\alpha_{i+2,i-1}^{(k)}$，形成相应的矩阵.

3.1.2.2）确定 Z_{i-1} 消去 $\beta_{i+2,i+1}^{(k)}$，$\beta_{i+2,i}^{(k)}$，形成相应的矩阵.

3.1.2.3）确定 Z'_{i-1} 消去 $\beta_{i+1,i}^{(k)}$，形成相应的矩阵.

3.1.3）NEXT i.

3.1.4）确定 P_{n-2} 消去 $\alpha_{n,n-2}^{(k)}$，形成相应的矩阵.

3.1.5）确定 Z_{n-2} 消去 $\beta_{n,n-1}^{(k)}$，形成相应的矩阵.

3.1.6）如果满足收敛准则，那末置 $\lambda_i=\alpha_{jj}^{(k)}/\beta_{jj}^{(k)}$ 并转4）.

3.2）NEXT k

4）NEXT j

5）END.

以上算法仅是一个简单的示意描述。许多具体细节，读者可进一步参阅 Moler，C. B. 和 Stewart G. W.的文章［9］。这里还应特别指出的是，和QR算法类似，在 QZ 迭代过程中，每迭代一次应由下到上地检查 A_k 的次对角元素。若某次对角元可忽略，则可将问题分裂为二个低阶的问题，从而分别进行迭代。

对于（4.2.3），其特征向量容易算出．事实上，经过有限次酉等价变换后迭代终止，则有

$$\hat{A} = \hat{U}\tilde{A}\tilde{V}, \tag{4.2.16}$$

$$\hat{B} = \hat{U}\hat{B}\tilde{V}. \tag{4.2.17}$$

这里$\hat{A}$与$\hat{B}$可以看作上三角阵（$\hat{A}$ 的次对角元相当小），而$\hat{U}$与$\hat{V}$为正交阵．据（4.2.16)和(4.2.17)有

$$\tilde{B}^{-1}\tilde{A}\hat{V} = \hat{V}\hat{B}^{-1}\hat{A}. \tag{4.2.18}$$

若设$\hat{B}^{-1}\hat{A}z_i = \lambda_i z_i$，这里$\lambda_i$为问题（4.1.1）的广义特征值，那末据（4.2.18）即有

$$\begin{aligned}\tilde{B}^{-1}\tilde{A}(\hat{V}z_i) &= \hat{V}\hat{B}^{-1}\tilde{A}z_i \\ &= \lambda_i(\hat{V}z_i).\end{aligned} \tag{4.2.19}$$

由此可见，$\hat{V}z_i$即为问题（4.2.3）对于λ_i的广义特征向量．而$\hat{B}^{-1}\hat{A}$的特征向量可由齐次方程

$$(\hat{A} - \lambda_i\hat{B})z_i = 0 \tag{4.2.20}$$

确定之．综上所述，问题（4.1.1）对应λ_i的广义特征向量为$V\hat{V}z_i$．顺便指出，QZ算法需二个$n\times n$辅助数组，因此当n很大时，存贮量较大．

§3 Peters-Wilkinson 方法

正如引言中指出那样，为了不破坏矩阵 **A**、**B** 的对称带状结构，**Peters** 和 **Wilkinson** 将标准对称特征值问题的 **Givens-Householder** 方法的思想用来解决对称广义特征值问题．为此，我们首先回顾一下标准对称特征值问题的 **Givens-Householder** 方法的基本依据：

（1）设**A**为对称阵，它具有相异的特征值，则其首主子矩阵行列式序列

$$P_0(\lambda),\ P_1(\lambda), \cdots, P_k(\lambda), \cdots, P_n(\lambda) \tag{4.3.1}$$

为Sturm序列，其中$P_0(\lambda) = 1$，$P_k(\lambda) = \det(A^{[k]} - \lambda I)$；

（2）在以上的假定下，记$a(\sigma)$为序列（4.3.1）在σ处相邻两多项式同号的数目，则矩阵A恰有$a(\sigma)$个特征值不小于σ。

现在设法将以上结论过渡到广义特征值问题。如§1中所说，对称广义特征值问题等价于对称标准特征值问题（4.1.6）。因此，如果矩阵C也具有相异的特征值，那末即可利用以上二个结论。当然，为了不破坏矩阵A、B的对称带状结构，我们并不希望真正去构成矩阵C，因此就不能直接利用序列$\{\det(C^{[k]}-\lambda I)\}$的同号数来隔离$(A-\lambda B)$的特征值。然而，$\det(A^{[k]}-\lambda B^{[k]})$与$\det(C^{[k]}-\lambda I)$ $(k=1, 2, \cdots, n)$在以上假设条件下，具有相同的符号，这样问题也就迎刃而解了。事实上，若设B的Cholesky分解为$B=LL^T$，则

$$L^{[k]}L^{[k]T}=B^{[k]},$$

$$(L^{[k]})^{-1}A^{[k]}(L^{[k]})^{-T}=C^{[k]};$$

于是

$$\begin{aligned}
&\det(A^{[k]}-\lambda B^{[k]})\\
&=\det[L^{[k]}\{(L^{[k]})^{-1}A^{[k]}(L^{[k]})^{-T}-\lambda I\}L^{[k]T}]\\
&=\det[L^{[k]}(C^{[k]}-\lambda I)L^{[k]T}]\\
&=[\det(L^{[k]})]^2\cdot\det(C^{[k]}-\lambda I).
\end{aligned}$$

因为$L^{[k]}(k=1, \cdots, n)$为非异，故$\det(A^{[k]}-\lambda B^{[k]})$的符号必与$\det(C^{[k]}-\lambda I)$ $(k=1, 2, \cdots, n)$的符号一致。根据以上事实即可得到如下结论：

在以上假设条件下，不小于σ的$A-\lambda B$特征值的数目恰为序列

$$\det(A^{[k]}-\sigma B^{[k]}) \qquad (k=0, 1, \cdots, n) \qquad (4.3.2)$$

在σ处的同号数$a(\sigma)$，这里规定$\det(A^{[0]}-\lambda B^{[0]})=1$。

但在这里应特别指出，序列（4.3.1）和序列（4.3.2）在计算上是有区别的。前者因为矩阵A可用初等镜象变换化约为对称三对角阵，所以序列（4.3.1）可以递推计算。可是序列（4.3.2）的计算却无此方便。

通常采用列主元素消去法计算$\det(A^{[k]}-\sigma B^{[k]})$的值，以保证计算的稳定性。但应注意，这时列主元素消去法是在k阶首主子矩阵中进行的，而k是从2逐次递增至n。所以，这里列主元素消去法与求解线性方程组时的情形略有不同。其算法简单描述如下：

算法 4.3.1 设给定对称矩阵A，$B\in R^{n\times n}$，B为正定，欲求$\det(A^{[k]}-\sigma B^{[k]})$的符号$s_k(k=1,2,\cdots,n)$，记$C=A-\sigma B$，$\sigma$为给定的数。

1）置$s_1=\mathrm{sign}\gamma_{11}$。

2）对$k=1,2,\cdots,n-1$

  2.1）置$s_{k+1}=s_k$。

  2.2）对$i=1,2,\cdots,k$

    2.2.1）若$|\gamma_{k+1,i}|>|\gamma_{ii}|$，则$\gamma_{k+1,j}\longleftrightarrow\gamma_{i,j}(j=i,i+1,\cdots,n)$，否则转2.2.3）。

    2.2.2）$\gamma_{k+1,i}$与γ_{ii}同号，则置$s_{k+1}=-s_{k+1}$。

    2.2.3）置$\gamma_{k+1,i}=\dfrac{\gamma_{k+1,i}}{\gamma_{ii}}$，

  2.2.4）置$\gamma_{k+1,j}=\gamma_{k+1,j}-\gamma_{k+1,i}\gamma_{i,j}(j=i+1,\cdots,n)$，

  2.3）NEXT i，

  2.4）如果$\gamma_{k+1,k+1}<0$，则置$S_{k+1}=-S_{k+1}$。

3）NEXT k。

4）END。

例4.3.1 利用算法4.3.1，确定对称矩阵

$$C=\begin{pmatrix}1&4&0&0\\4&2&1&0\\0&1&3&4\\0&0&4&3\end{pmatrix}$$

各阶首主子式的符号。

解 显然$s_0=1$，$s_1=1$。按语句（2.2.1—2.2.4）执行如下：$k=1$时变换为：

$$C\xrightarrow[\text{二行交换}]{\text{第一行与第}}\begin{pmatrix}4&2&1&0\\1&4&0&0\\0&1&3&4\\0&0&4&3\end{pmatrix}\xrightarrow[\frac{1}{4}\times\text{第一行}]{\text{第二行减去}}\begin{pmatrix}4&2&1&0\\0&7/2&-1/4&0\\0&1&3&4\\0&0&4&3\end{pmatrix}.$$

因为进行过一次行交换且γ_{11}与γ_{21}同号，同时$\gamma_{22}>0$，故$s_2=-1$。$k=2$时变换为

$$\xrightarrow[\frac{2}{7}\times\text{第二行}]{\text{第三行减去}}\begin{pmatrix}4&2&1&0\\0&7/2&-1/4&0\\0&0&43/14&4\\0&0&4&3\end{pmatrix}.$$

因为没有进行过行交换，而$43/14>0$，故据语句2.1）知$s_3=-1$。$k=3$时变换为

$$\xrightarrow[\text{四行交换}]{\text{第三行与第}}\begin{pmatrix}4&2&1&0\\0&7/2&-1/4&0\\0&0&4&3\\0&0&43/14&4\end{pmatrix}\xrightarrow[\text{去 }43/56\times\text{第三行}]{\text{第四行减}}\begin{pmatrix}4&2&1&0\\0&7/2&-1/4&0\\0&0&4&3\\0&0&0&95/56\end{pmatrix}.$$

因为进行过一次交换，γ_{33}与γ_{43}同号而$95/56>0$，所以$s_4=1$。

从以上简单的例子可以看出，P-W 算法的主要优点是：能够充分利用原始矩阵A和B的稀疏带状的特点，从而省去了许多存贮量与计算量。因此，它适用于求解大型对称广义特征值问题。事实上，若记$A-\sigma B$的半带宽为$m+1$（包括对角元），则算法4.3.1的第k步仅涉及第$k-m+1$到第$k+1$行的带内元素。尽管由于行交换，原来矩阵的半带宽可能要变大，但最大不会超过$2m+1$。

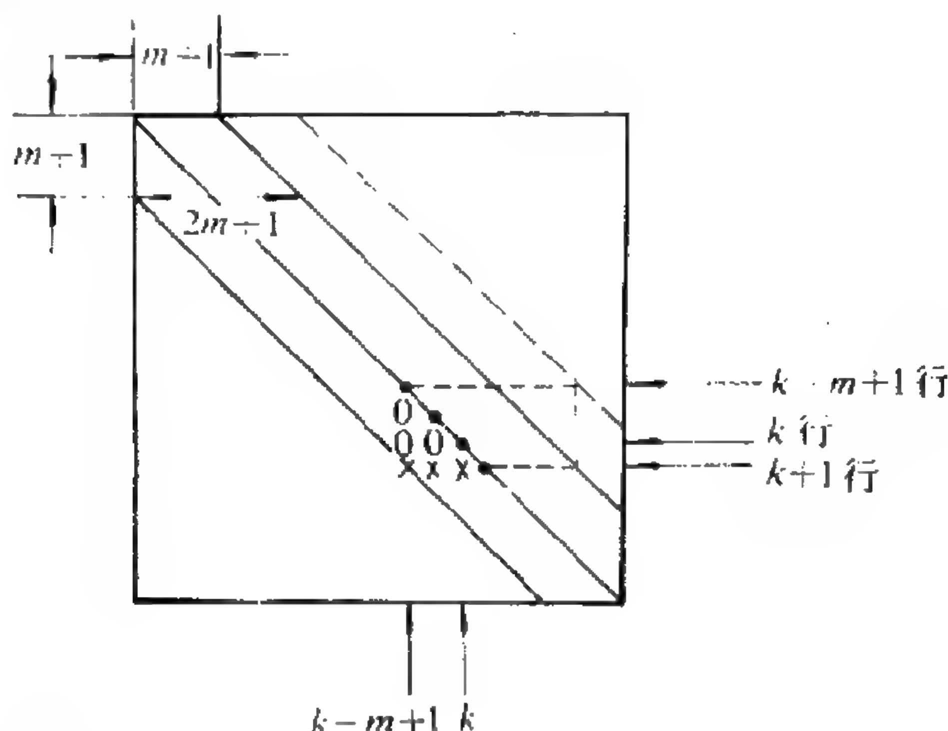


图 4.1 A-σB在列主元消去过程中带宽变化

P-W算法的结构基本上和 Givens-Householder 算法是一致的，即利用算法4.3.1确定序列(4.3.2)的同符号数，然后用二分法来隔离特征值（部分或全部）。二分法可以与求特征向量的反迭代法以及Rayleigh商迭代结合起来，相辅相成。二分法不必进行到隔离出精度范围的特征值为止，而只需进行到将特征值隔离开为止。然后，用反迭代求近似特征向量，再利用Rayleigh 商使特征值精确化，如此往复循环，直至近似特征值符合精度标准为止。

因为当B为非异时，广义特征值问题（4.1.1）与标准特征值问题（4.1.3）是等价的*。而后者的反迭代格式是：

$$(B^{-1}A-\tilde{\lambda}_kI)\boldsymbol{x}_k^{(i+1)}=\boldsymbol{x}_k^{(i)}\tau_i \qquad (i=0,1,\cdots) \quad (4.3.3)$$

其中$\sigma_k^{(i)}$为位移量，而τ_i为规范化常数。（4.3.3）式两边左乘以B，即得

$$(A-\tilde{\lambda}_kB)\boldsymbol{x}_k^{(i+1)}=B\boldsymbol{x}_k^{(i)}\tau_i \qquad (i=0,1,\cdots), \quad (4.3.4)$$

这就是广义特征值问题$A-\lambda B$的反迭代格式。和标准特征值问题一样，由方程组（4.3.4）求得$X_k^{(i+1)}$后，利用Reyleigh商

$$\rho(\boldsymbol{x}_k^{(i+1)})=\frac{\boldsymbol{x}_k^{(i+1)T}A\boldsymbol{x}_k^{(i+1)}}{\boldsymbol{x}_k^{(i+1)T}B\boldsymbol{x}_k^{(i+1)}}, \quad (4.3.5)$$

将特征值精确化最为合适（参阅下节及习题4.6）。根据以上分析即可得到下面的算法。

算法4.3.2 设A、$B\in R^{n\times n}$且B为正定，又假定广义特征值问题$A-\lambda B$具有相异的特征值，求$A-\lambda B$第k个特征值及其对应的特征向量。

1）设σ为λ_k的一个下界，利用算法4.1.2计算序列（4.3.2）的符号，以同符号数$a(\sigma)$确定$[\sigma,\infty)$中$A-\lambda B$的特征值的个数（NUMBER子程序）；

2）用二分法将$A-\lambda B$的第k个特征值隔离在区间$[\lambda_k^-,\lambda_k^+]$内，取初始特征值为$\lambda_k^{(0)}=\frac{1}{2}(\lambda_k^-+\lambda_k^+)$或利用割线法确定$\lambda_k^{(0)}$（BISECT和INTPOL子程序）；

3）选取初始特征向量$\boldsymbol{x}_k^{(0)}$。通常的选取法是，用部分主元消去将$(A-\lambda_k^{(0)}B)$化为上三角阵U，记下所作的行交换。然后求解三角形方程组

*）这里当然亦可将对称广义特征值问题化为对称标准特征值问题，进而考虑它的反迭代格式。

$$Ux_k^{(0)}=e.\qquad(4.3.6)$$

这里 e 是分量全为 1 的向量；

4）对 $i=0,1,\cdots$ （INVERS 子程序），

4.1）将 $x_k^{(i)}$ 规范化，得 $\tilde{x}_k^{(i)}$.

4.2）利用 $(A-\lambda_k^{(0)}B)$ 的 LU 分解及行交换信息求解线性方程组

$$(A-\lambda_k^{(0)}B)x_k^{(i+1)}=B\tilde{x}_k^{(i)},$$

4.3）利用 **Rayleigh** 商使 $\lambda_k^{(i)}$ 精确化

$$\lambda_k^{(i+1)}=\frac{x_k^{(i+1)T}Ax_k^{(i+1)}}{x_k^{(i+1)T}Bx_k^{(i+1)}},$$

4.4） 收敛判别：如果

$$\frac{|\lambda_k^{(i+1)}-\lambda_k^{(i)}|}{|\lambda_k^{(i+1)}|}\leqslant\varepsilon.$$

（这里 ε 是给定的精度）时，$\lambda_k^{(i+1)}$ 与 $x_k^{(i+1)}$ 即为所求的结果，转 6）.

5）**NEXT** i;

6）置 $\lambda_k=\lambda_k^{(i+1)}$，$x_k=x_k^{(i+1)}$；

7）**END**.

§4 子空间迭代法

乘幂法和反乘幂法(即反迭代法)是计算标准特征值问题的二个有效算法. 它们特别适应大型稀疏矩阵的特征问题，但它们只能求得占优特征值和特征向量. 然而实际问题往往需要求前几个占优特征值及其对应的特征向量，如多元统计分析中的主成分分析等. 过去常采用降阶法，逐一求出前几个占优特征

值和特征向量。但是，降阶法对各种不同结构的矩阵并无统一的算法；另外，降阶还将破坏原矩阵的形状以及受到舍入误差较大的影响。

子空间迭代法在一定程度上能克服降阶法这些缺点。其基本思想是：对若干个试验向量（可看作 $\boldsymbol{R}^n$ 中某子空间的基）同时进行迭代，从而同时得到前几个占优特征值及其对应的特征向量（或者说 A 的占优子空间）。所以，子空间迭代法可看成幂法、反幂法的一种推广。

这里我们仅讨论实对称阵的情形，至于非对称阵的子空间迭代法，读者可参阅 **Stewart, G. W.** 以及 **Cline, M., Jennings, A.** 等的文章[10]、[11]。为说明方便起见，设 $A\in\boldsymbol{R}^{n\times n}$ 为对称正定阵而

$$0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$$

和

$$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$$

分别为它的特征值和对应的正交规范特征向量。现欲求 A 的 p 个占优特征对 $(\lambda_i,\ \boldsymbol{x}_i)\,(i=1,2,\cdots,p)$。

设 $Q_0=(\boldsymbol{q}_1^{(0)},\boldsymbol{q}_2^{(0)},\cdots,\boldsymbol{q}_p^{(0)})$，而 $Q_0^TQ_0=I_p$。$\boldsymbol{q}_1^{(0)}$ $\boldsymbol{q}_2^{(0)},\cdots,\boldsymbol{q}_p^{(0)}$ 可看作 p 个占优特征向量的初始近似，也可以看作一 p 维子空间的正交基。按幂法的思想，应以矩阵 A 同时左乘这些向量，得

$$V_0=AQ_0. \tag{4.4.1}$$

但是就一般而言，矩阵 V_0 不再是列正交了，所以应对 V_0 的列正交化，这可以利用修正的 **Gram-Schmidt** 正交化算法来实现。于是得

$$V_0=Q_1R_1 \tag{4.4.2}$$

或

$$AQ_0=Q_1R_1. \tag{4.4.3}$$

这里$Q_1 \in \boldsymbol{R}^{n\times p}$为列正交阵，$R \in \boldsymbol{R}^{p\times p}$为上三角阵。进而又可取$Q_1$作为新的迭代矩阵，因为$Q_1$的列实际上是$V_0$列空间的正交基（注意$R_1$为非异）。一般的迭代格式为：

$$\begin{cases} V_k = AQ_k, \\ V_k = Q_{k+1}R_{k+1} \qquad (k=0,1,\cdots). \end{cases} \tag{4.4.4}$$

以上迭代格式常称为简单子空间迭代法。容易验证：

$$A^k Q_0 = Q_k R_k R_{k-1} \cdots R_1. \tag{4.4.5}$$

可以证明：当$k\to\infty$时，Q_k的列将趋于A的p个占优正交规范特征向量，但收敛速度很慢。

Q_k的列收敛速度缓慢的原因在于：在子空间$\mathrm{Span}(Q_k)$中，Q_k的列未必是A的p个占优特征向量在空间$\mathrm{Span}(Q_k)$中的最优近似。于是，我们必须研究以下问题：设给定$Q \in \boldsymbol{R}^{n\times p}$且$Q^TQ = I_p$，据此构成$p$维子空间$\mathfrak{S}^p \equiv \mathrm{Span}(Q)$。那末，在该子空间范围中，$A$的$p$个占优特征对的最优近似是什么？

如果$p=1$，即给定向量q且$\|q\|_2=1$，它作为A的近似特征向量，对应的A的近似特征值γ应取何值为最优呢？很自然，如果γ值使残量$\|Aq-q\gamma\|_2$为极小，那末该γ值即为最优值。据广义的勾股定理知最优的γ使

$$r(q) = Aq - q\gamma \perp q$$

亦即

$$q^T(Aq - q\gamma) = 0$$

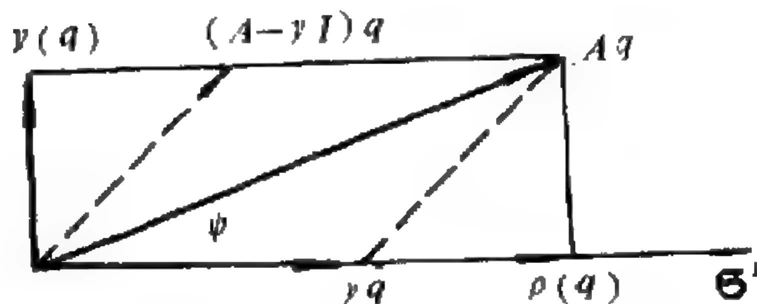


图 4.2 广义勾股定理

由此即可推知最优解即为 A 的 Rayleigh 商

$$\gamma = \gamma q^T q = q^T A q = \rho(q). \qquad (4.4.6)$$

如果$p>1$，则有以下定理

定理4.4.1 设$A\in R^{n\times n}$为对称阵，给定 $Q\in R^{n\times p}$且$Q^TQ=I_p$，则对任何$B\in R^{p\times p}$成立不等式

$$\|AQ-QH\|_p \leqslant \|AQ-QB\|_p \quad (p=2,F), \qquad (4.4.7)$$

其中 $H=Q^TAQ=\rho(Q)$。

证 因为

$$(AQ-QB)^T(AQ-QB)=Q^TA^2Q-H^2+(H-B)^T(H-B),$$

而

$$\|AQ-QB\|_F^2 = \mathrm{tr}[(AQ-QB)^T(AQ-QB)],$$

所以

$$\|AQ-QB\|_F^2 \geqslant \|AQ-QH\|_F^2.$$

另一方面，因为$(H-B)^T(H-B)$为非负定，故

$$\begin{aligned}\|AQ-QB\|_2^2 &= \max[\lambda((AQ-QB)^T(AQ-QB))]\\ &\geqslant \max[\lambda((AQ-QH)^T(AQ-QH))]\\ &= \|AQ-QH\|_2^2.\end{aligned}$$

以上定理说明，虽然 $\mathfrak{S}^p\equiv\mathrm{span}(Q)$不是$A$ 的不变子空间，可是当取B为 A 的 Rayleigh 商矩阵 $\rho(Q)$时，能使残量$\|AQ-QB\|_p(p=2,F)$极小。

现设 $\lambda(H)$：$\theta_1\leqslant\theta_2\leqslant\cdots\leqslant\theta_p$，那末它们是不是由子空间$\mathfrak{S}^p\equiv\mathrm{span}(Q)$推得的$A$的$p$个占优特征值的最优近似呢？

如果$\mathfrak{S}^p$为$A p$个占优特征向量张成的不变子空间，则显然有

$$\lambda_{n-p+j}=\min_{\mathfrak{S}^j\subset\mathfrak{S}^p}\ \max_{\substack{g\in\mathfrak{S}^j\\ g\neq 0}} \frac{g^TAg}{g^Tg} \quad (j=1,\cdots,p),$$

其中$\mathfrak{S}^j$为$\mathfrak{S}^p$中任意j维子空间。如果$\mathfrak{S}^p$不是A的p 个占优特征

向量张成的不变子空间，则很自然地定义：

$$\beta_j = \min_{\mathfrak{S}^j \subset \mathfrak{S}^p} \max_{\substack{g \in \mathfrak{S}^j \\ g \neq 0}} \frac{\mathbf{g}^T A \mathbf{g}}{\mathbf{g}^T \mathbf{g}} \quad (j = 1, \cdots, p). \qquad (4.4.8)$$

把它作为由$\mathfrak{S}^p$推得的λ_{n-p+j}的最优近似是合理的。那末，β_j与θ_j之间有什么关系呢？以下定理回答了这个问题。在证明定理之前，首先应注意子空间之间存在下列关系：

$$\begin{array}{lll}
\boldsymbol{R}^n & & \\
\cup & & \\
\mathfrak{S}^p \equiv Q\boldsymbol{R}^p & & (Q \in \boldsymbol{R}^{n \times p}) \\
\cup \qquad \cup & & \\
\mathfrak{S}^j \equiv Q\mathscr{T}^j \equiv QG\boldsymbol{R}^j & & (G \in \boldsymbol{R}^{p \times j}).
\end{array}$$

请注意，$\mathfrak{S}^j$为$\mathfrak{S}^p$中的j维子空间，也是$\boldsymbol{R}^n$中的j维子空间，而$\mathscr{T}^j$为$\boldsymbol{R}^p$中的j维子空间。由以上关系可见，$\mathfrak{S}^j \subset \boldsymbol{R}^n$与$\mathscr{T}^j \subset \boldsymbol{R}^p$之间存在一一对应的关系。

定理 4.4.2 $\beta_j = \lambda_j(H) \equiv \theta_j \qquad (j = 1, \cdots, p)$

证
$$\beta_j = \min_{\mathfrak{S}^j \subset \mathfrak{S}^p} \max_{\substack{g \in \mathfrak{S}^j \\ g \neq 0}} \frac{\mathbf{g}^T A \mathbf{g}}{\mathbf{g}^T \mathbf{g}} \quad (\mathbf{g} = Q\mathbf{s},\ \mathbf{s} \in \mathscr{T}^j,\ \mathbf{g}^T\mathbf{g} = \mathbf{s}^T Q^T Q \mathbf{s} = \mathbf{s}^T \mathbf{s} \neq 0)$$

$$= \min_{\mathscr{T}^j \subset R^p} \max_{\substack{s \in \mathscr{T}^j \\ s \neq 0}} \frac{\mathbf{s}^T Q^T A Q \mathbf{s}}{\mathbf{s}^T \mathbf{s}}$$

$$= \min_{\mathscr{T}^j \subset R^p} \max_{\substack{s \in \mathscr{T}^j \\ s \neq 0}} \frac{\mathbf{s}^T H \mathbf{s}}{\mathbf{s}^T \mathbf{s}}$$

$$= \theta_j.$$

由以上定理可见：H的特征值确是由子空间$\mathfrak{S}^p$推得的$\mathbf{A}$的p个占优特征值的最优近似。

设$H\mathbf{g}_j = \theta_j \mathbf{g}_j \quad (j = 1, \cdots, p)$，$\|\mathbf{g}\|_2 = 1$且$\mathbf{g}_j^T \mathbf{g}_i = 0 \ (i \neq j)$，

则

$$\begin{pmatrix} \theta_1 & & 0 \\ & \ddots & \\ 0 & & \theta_p \end{pmatrix} = G^T HG = (QG)^T A(QG). \tag{4.1.9}$$

这里$G=(g_1, \cdots, g_p)$。因为θ_i可视为λ_{n-p+i}的最优近似，所以取$y_i = Qg_i$ 作为对应特征向量的最优近似也是合理的，进一步的论述可参见本章习题 4.8。θ_i 和 y_i 分别称为矩阵 A 在子空间 Span(Q) 上的Ritz值和 Ritz 向量。Ritz值和 Ritz 向量总称为矩阵A在子空间 Span(Q) 上的 Rayleigh-Ritz 近似。

综上所述即可给出 Rayleigh-Ritz 过程，以下简称为 RR 过程。

RR 过程给定对称阵 $A\in R^{n\times n}$ 及 p 维子空间$\mathfrak{S}^p=\{Sx|x\in R^p\}$，其中 $S\in R^{n\times p}$ 而 rank(S) $=p$。利用$\mathfrak{S}^p$欲求A的p个占优特征对的最优近似(θ_i, y_i) $(i=1, \cdots, p)$。

(i) 调用修正的Gram-Schmidt 子程序，正交规范化S 的列即

$$S=QR;$$

(ii) 调用子程序OP构成 AQ;

(iii) 构成Q的 Rayligh 商

$$H=\rho(Q)=Q^T(AQ);$$

(iv) 调用Jacobi 子程序，作谱分解

$$Hg_i = g_i\theta_i \qquad (i=1, \cdots, p);$$

(v) 形成 Ritz 向量

$$y_i = Qg_i \qquad (i=1, \cdots, p).$$

上述过程中OP子程序的功能是：矩阵 A 毋需用一个 $n\times n$ 数组给出，而给出了对任意向量v计算 $A*v$ 的规则。在该子程序中，矩阵 A 的存贮以 $A*v$ 的计算都充分利用A的稀疏性。

至此，我们即可将简单子空间迭代法和 Rayleigh-Ritz 过程结合起来，产生新的子空间迭代算法。

算法 4.4.1 设给定对称阵 $A \in \boldsymbol{R}^{n \times n}$ 及 $S_0 \in \boldsymbol{R}^{n \times p}$，而 rank$(S_0) = p$，从子空间 $\mathfrak{S} = \text{span}(S_0)$ 出发，求 A 的 p 个占优特征对。

1）对 $k = 1, 2, \cdots,$

1.1）调用 OP 子程序形成

$$C_k = AS_{k-1}.$$

1.2）调用修正的 Gram-Schmidt 子程序，正交规范化 C_k 的列，即

$$C_k = Q_k R_k,$$

1.3）调用 OP 子程序构成 AQ_k，

1.4）构成 Rayleigh 商矩阵

$$\hat{H}_k = Q_k^T A Q_k,$$

1.5）调用 Jacobi 子程序，谱分解

$$\hat{H}_k = G_k \Theta G_k^T, \quad (\Theta = \text{diag}(\theta_1, \cdots, \theta_p)),$$

1.6）形成 Ritz 向量

$$S_k = Q_k G_k,$$

1.7）检验是否达到收敛标准，若已达到收敛标准即转3）。

2）NEXT k。

3）输出 $(\theta_i, \boldsymbol{y}_i)$ $(i = 1, \cdots, p)$。

4）END.

一般而言，算法 4.4.1 中 G_k 为一系列正交矩阵的乘积，设

$$G_k = p_1 \cdots p_l$$

那末语句1.5）和语句1.6）可同时进行如下：

1）置$S_k = Q_k$；

2）对 $i = 1, \cdots, t$，置 $S_k = S_k p_i$。

这样矩阵G_k就毋需明显形成，在整个计算过程中只要一个$n \times p$数组就是以应付S、C 以及Q的计算了。其次，读者应注意，在算法4.4.1中，S_k 的列是子空间 $A^k\mathfrak{S}$ 的基，C_k 的列也是子空间 $A^k\mathfrak{S}$ 的基。但前者应较后者收敛于A的 p个占优特征向量要快．当然迭代一次的计算量相对而言较高。特别是当矩阵阶数较高时，在执行语句 1.3)时要额外调用 OP 子程序 p 次，显然算法4.4.1的计算量要大大超过简单子空间迭代法的计算量。为此，以下我们来讨论如何避免额外调用p次 OP子程序。

因为考虑到A^{-2}与 A 具有相同的特征向量，所以也可针对 A^{-2} 来实施 Rayleigh-Ritz 过程。

如同算法4.4.1，首先计算 $C_k = AS_{k-1}$，然后设法形成新基C_kF_k，这里 $F_k \in R^{p \times p}$应满足以下二个条件：

① $(C_kF_k)^T(C_kF_k) = I_p$，　　　　（正交性）。

② $(C_kF_k)^T A^{-2}(C_kF_k) = D_k^{-2}$，　　　　（Ritz 向量性质）．

其中$D_k = \mathrm{diag}(\delta_1, \cdots, \delta_p)$为 p 阶对角阵。②式亦可写成

$$(F_k^T S_{k-1}^T A) A^{-2} (AS_{k-1}F_k) = F_k^T F_k = D_k^{-2}. \quad (4.4.10)$$

由此可见，F_kD_k 为 正交阵。这样据①式及(4.4.10)即有

$$C_k^T C_k = F_k^{-T} F_k^{-1} = (F_k^{-T} D_k^{-1}) D_k^2 (D_k^{-1} F_k^{-1})$$

$$= (F_k D_k) D_k^2 (F_k D_k)^T. \quad (4.4.11)$$

由此可见，F_k和D_k即可由$C_k^T C_k$的谱分解确定。

算法4.4.2　设给定对称阵$A \in R^{n \times n}$及$S_0 \in R^{n \times p}$而$\mathrm{rank}(S_0) = p$。从子空间 $\mathfrak{S} = \mathrm{span}(S_0)$出发，欲求$A$的$p$个占优特征对。

1）对$k = 1, 2, \cdots$，

1.1）调用 OP 子程序形成

$$C_k = AS_{k-1},$$

1，2）构成矩阵

$$\dot{H}_k = C_k^T C_k,$$

1，3）调用Jacobi 子程序，作谱分解

$$\dot{H}_k = B_k D_k^2 B_k^T,$$

1，4）形成

$$S_k = C_k B_k D_k^{-1},$$

1，5）检验是否达到收敛标准，若已达到收敛标准即转3）。

2）**NEXT** k。

3）输出(d_i, y_i) $(i=1, \cdots, p)$，这里 y_i 是 S_k 的第 i 列。

4）**END**。

应注意算法4.4.2中的S_k 和算法 4.4.1 中的 S_k 是不同的，而且算法 4.4.2 中的S_k 并不能表成一系列正交阵的乘积。为此对算法 4.4.2 仍须作进一步修改。考虑到

$$\dot{H}_k = C_k^T C_k = R_k^T Q_k^T Q_k R_k = R_k^T R_k = L_k R_k. \qquad (4.4.12)$$

如果记 $H_k \equiv R_k L_k$，那末据 LR 算法知，H_k 较 $\dot{H}_k$ 更接近对角阵。但据(4.4.12)应有

$$\begin{aligned} H_k &\equiv R_k L_k \equiv R_k R_k^T = R_k \dot{H}_k R_k^{-1} \\ &= R_k B_k D_k^2 B_k^T R_k^{-1} \\ &= (R_k B_k D_k^{-1}) D_k^2 (D_k B_k^T R_k^{-1}). \end{aligned} \qquad (4.4.13)$$

由上式知，H_k与$\dot{H}_k$相似，所以用 H_k 替代 $\dot{H}_k$ 是合理的。其次，不难验证，$P_k \equiv R_k B_k D_k^{-1}$是$H_k$的正交特征矩阵，于是算法4.4.2中的$S_k$即可表为

$$S_k \equiv C_k B_k D_k^{-1} = Q_k R_k B_k D_k^{-1} = Q_k P_k.$$

这样S_k就有可能表成一系列正交阵的乘积了。综上所述，即可构成如下 RITZIT 算法。

算法4.4.3 设给定对称阵 $A \in R^{n\times n}$ 及 $S_0 \in R^{n\times p}$，而

$\text{rank}(S_0)=p$，从子空间$\mathfrak{S}=\text{span}(S_0)$出发，求$A$的$p$个占优特征对。

1)对$k=1,2,\cdots$,

1.1)调用 OP 子程序形成

$$C_k = AS_{k-1},$$

1.2)调用修正的Gram-Schmidt 子程序，正交规范化 C_k 的列，即

$$C_k = Q_k R_k,$$

1.3)形成矩阵

$$H_k = R_k R_k^T,$$

1.4)调用Jacobi 子程序，作谱分解

$$H_k = P_k D_k^2 P_k^T,$$

1.5)形成 Ritz 向量

$$S_k = Q_k P_k,$$

1.6)检验是否达到收敛标准，若已达到收敛标准，即转3)。

2)NEXT k。

3)输出$(\delta_i, \ y_i)$　$(i=1,\cdots,p)$。

4)END。

在一定的假设下，可以证明子空间迭代法的收敛性。我们不加证明引入以下定理。

定理4.4.3　设$A\in \boldsymbol{R}^{n\times n}$为对称正定阵，其特征值及其对应的特征向量分别为

$$0<\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_{n-p}<\lambda_{n-p+1}\leqslant\cdots\leqslant\lambda_n$$

和

$$z_1, \ z_2, \cdots, z_n.$$

设$S\in\boldsymbol{R}^{n\times p}$且$S^TS=I_p$，若$Z^TS$为非异，则由算法4.4.3产生的 Ritz 向量当$k\to\infty$时有

$$\sin(y_i^{(k)}, z_{n-p+i}) = O\left[\left(\frac{\lambda_{n-p}}{\lambda_{n-p+i}}\right)^k\right], \tag{4.4.14}$$

$$(i = 1, \cdots, p)$$

这里$Z = (z_n,\ z_{n-1}, \cdots, z_p)$。

据以上定理知，当 $i = p$ 时有

$$\sin(y_p^{(k)}, z_n) = O\left[\left(\frac{\lambda_{n-p}}{\lambda_n}\right)^k\right]. \tag{4.4.15}$$

而一般而言，$\left|\frac{\lambda_{n-p}}{\lambda_n}\right| < \left|\frac{\lambda_{n-1}}{\lambda_n}\right|$，所以利用RR过程等措施改进后的子空间迭代法，显然要比简单子空间迭代法有效得多。

由(4.4.14)还可以看出：在相同的迭代次数前提下，愈是排在后面的特征值及其对应的特征向量，它们的计算精度愈是高。因此，在实际计算时，往往把p取得比实际所求的个数r稍大一点。有的文献推荐，取$p = \min(2r,\ r+8)$，这额外的$p-r$个向量称为护卫向量。显然，对于固定的r，p取得愈大，迭代次数就愈小。但这样每次迭代所花费的工作量就较大。因此这里有个权衡得失的问题，这需在实际计算中去探索。至于S_0的选取，可以先取 p 个由随机数构成的向量，然后将它们正交规范化。

最后我们指出，如果需要求A 前面 p 个按模小的特征值，只要在以上算法中将 A 换成A^{-1}即可。

下面我们来讨论对称广义特征值问题的子空间迭代法。正如§1指出那样，对称广义特征值问题

$$Ax = \lambda Bx \tag{4.4.16}$$

可化为标准特征值问题

$$\begin{aligned} &\tilde{A}y = \lambda y, \\ &\tilde{A} = L^{-1}AL^{-T}, \\ &y = L^T x, \\ &B = LL^T. \end{aligned} \tag{4.4.17}$$

于是算法 4.4.3 和(4.4.17)相结合，即可得到对称广义特征值问题的子空间迭代算法。

现在我们来分析，用子空间迭代法求解对称广义特征值问题的独到之处。对于对称广义特征值问题而言，算法 4.4.3 中语句1.1)实际上应执行

$$C_k = L^{-1}AL^{-T}S_{k-1}, \tag{4.4.18}$$

而

$$S_{k-1} = L^T x_{k-1}. \tag{4.4.19}$$

实际上，以上运算可以分以下二步来完成：

(i) 调用 OP 子程序计算

$$Z_k = Ax_{k-1}; \tag{4.4.20}$$

(ii) 用向后回代法求解

$$LC_k = Z_k. \tag{4.4.21}$$

如果 A，B 都是稀疏矩阵，那末在执行以上几步运算时，就可充分利用它们的稀疏性质。首先，在执行迭代之前，对 B 作 Cholesky 分解，这可利用后面第六章所介绍的对称变带宽矩阵 Cholesky 分解方法。求得 L 后，在迭代过程中只要反复求解(4.4.21)形状的方程组集，每次迭代时仅Z_k在变化。其次，对矩阵A亦可利用第六章介绍的任何稀疏矩阵压缩存贮技巧，以提高 OP 子程序的效率。

算法4.4.4　设给定对称广义特征值问题 $Ax = \lambda Bx$，设 $S_0 \in R^{n\times p}$且 $S_0^T S_0 = I_p$，求问题的前 p 个占优特征值及其对应的特征向量。

1) 对 B 作Cholesky分解

$$B = LL^T.$$

2) A 按某种对称稀疏阵存贮方法存贮

3) 求解

$$L^T X_0 = S_0.$$

4) 对 $k=1, 2, \cdots$,

4.1) 调用 OP 子程序计算

$$Z_k = AX_{k-1},$$

4.2) 求解

$$LC_k = Z_k,$$

4.3) 调用修正的 Gram-schmidt 子程序，作

$$C_k = Q_k R_k,$$

4.4) 形成

$$H_k = R_k R_k^T,$$

4.5) 调用 Jacobi 子程序，作谱分解

$$H_k = P_k D_k^2 P_k^T,$$

4.6) 形成 $\quad S_k = Q_k P_k,$

4.7) 求解

$$L^T X_k = S_k,$$

4.8) 检验是否达到收敛标准，若已达到收敛标准即转 6)。

5) NEXT k.

6) 输出 $(\delta_i, x_i) \quad (i=1, \cdots, p)$.

7) END

第四章习题

4.1 求解以下广义特征值问题 $Ax = \lambda Bx$:

(i) $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$;

(ii) $A = \begin{pmatrix} 2 & 2 \\ 2 & 4 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 \\ 2 & 3 \end{pmatrix}$;

$$(iii)\ A=\begin{pmatrix}5 & -2 & 0\\ -2 & 3 & -1\\ 0 & -1 & 1\end{pmatrix},\qquad B=\begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 3\end{pmatrix}.$$

4.2 将以下对称广义特征值问题 $Ax=\lambda Bx$ 化为标准特征值问题(4.1.6)：

$$(i)\ A=\begin{pmatrix}1 & 2\\ 2 & 4\end{pmatrix},\qquad B=\begin{pmatrix}2 & 1\\ 1 & 4\end{pmatrix};$$

$$(ii)\ A=\begin{pmatrix}-2 & 1 & 0 & 0\\ 1 & -2 & 1 & 0\\ 0 & 1 & -2 & 1\\ 0 & 0 & 1 & -2\end{pmatrix},\quad B=\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 4 & 0 & 0\\ 0 & 0 & 4 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}.$$

4.3 设 A，$B\in R^{n\times n}$ 且 B 为非异，试证存在酉阵 U 和 V，使 A，B 同时酉等价于上三角阵 $\widetilde{A}$，$\widetilde{B}$：

$$\widetilde{A}=U^HAV\qquad \widetilde{B}=U^HBV.$$

4.4 设 A，$B\in R^{n\times n}$ 为对称矩阵且 B 为正定，$Ax=\lambda Bx$ 的特征值为 $\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$，试证：

$$\lambda_j=\min_{\substack{\dim(\mathscr{T})\\=j}}\max_{\substack{x\in\mathscr{T}\\x\neq 0}}\frac{x^TAx}{x^TBx}\qquad (j=1\cdots,\ n).$$

这里 $\mathscr{T}$ 为 R^n 中的 j 维子空间。

4.5 设 A，B，$C\in R^{n\times n}$ 皆为对称正定，试证二次特征值问题：

$$(A-\lambda B-\lambda^2C)x=0.$$

(i) 等价于标准特征值问题：

$$(G-\lambda I_{2n})y=0$$

其中

$$G=\begin{pmatrix}O & I_n\\ C^{-1}A & -C^{-1}B\end{pmatrix},\quad y=\begin{pmatrix}x\\ \hat{x}\end{pmatrix},\quad \hat{x}=\lambda x.$$

(ii) 等价于对称广义特征值问题

$$(E-\lambda F)y=0$$

其中

$$E=\begin{pmatrix} C & O \\ O & A \end{pmatrix},\quad F=\begin{pmatrix} O & C \\ C & B \end{pmatrix},\quad y=\begin{pmatrix} \hat{x} \\ x \end{pmatrix},\quad \hat{x}=\lambda x.$$

4.6 设对称特征值问题 $A-\lambda B$，它对应的 Rayleigh 商为

$$\rho(x)=\frac{x^T A x}{x^T B x},\quad x\neq 0.$$

试证：

(i) $\rho(\alpha x)=\rho(x)$，$\alpha\neq 0$。

(ii) $\lambda_1\leqslant\rho(x)\leqslant\lambda_n \qquad (\forall \|x\|_2=1)$，

(iii) $\nabla\rho(x)=2(Ax-\rho(x)Bx)^T/x^T Bx$，

(iv) $\|(A-\sigma B)x\|^2_{A^{-1}}\geqslant\|Ax\|^2_{A^{-1}}-|\rho(x)|^2\|Bx\|^2_{A^{-1}}$，

等号成立当且仅当 $\sigma=\rho(x)$。

4.7 设 $p=F$，试证(4.4.7)式等号成立当且仅当 $B=H$；设 $p=2$，试构造矩阵 A，Q 以及二阶阵 B，使 $B\neq H$ 但(4.4.7)式等号成立。

4.8 设 $A\in R^{n\times n}$ 为对称阵，$\mathscr{T}^p$ 为 R^n 中的 p 维子空间，试证对任何 $x\in\mathscr{T}^p$，$Ax-\rho(x)x\perp\mathscr{T}^p$ 当且仅当 x 为矩阵 A 在子空间 $\mathscr{T}^p$ 上的 Ritz 向量。

4.9 利用算法 4.3.1 确定以下对称矩阵：

$$\text{(i)}\ \begin{pmatrix} 1 & 0 & 1 & 2 \\ 0 & -1 & -2 & 0 \\ 1 & -2 & -2 & 3 \\ 2 & 0 & 3 & 6 \end{pmatrix},\quad \text{(ii)}\ \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & -1 & -2 & 0 \\ 1 & -2 & -2 & 3 \\ 2 & 0 & 3 & 6 \end{pmatrix}$$

各阶首主子式的符号。

4.10 编制算法4.4.3的电算程序，利用该程序计算矩阵

$$A=\begin{pmatrix} -1 & 0 & 3 & 0 \\ 0 & -1 & 0 & 0 \\ 3 & 0 & -1 & 0 \\ 0 & 0 & 0 & -3 \end{pmatrix}$$

的前二个占优特征对，精度自定。

4.11 编制算法 4.4.4 的电算程序，利用该程序计算 $Ax=\lambda Bx$ 的前二个占优特征对，精度自定。矩阵 A, B 为

$$A=\begin{pmatrix}10 & 2 & 3 & 1 & 11\\ 2 & 12 & 1 & 2 & 1\\ 3 & 1 & 11 & 1 & -1\\ 1 & 2 & 1 & 9 & 1\\ 1 & 1 & -1 & 1 & 15\end{pmatrix},\quad B=\begin{pmatrix}12 & 1 & -1 & 2 & 1\\ 1 & 14 & 1 & -1 & 1\\ -1 & 1 & 16 & -1 & 1\\ 2 & -1 & -1 & 12 & -1\\ 1 & 1 & 1 & -1 & 11\end{pmatrix}.$$

参 考 文 献

1. Stewart, G. W, (1973), *Introdction to Matrix Computation*, New York, Academic Press.
2. Jennings, A. (1977), *Matrix Computation for Engineers and Scientists*, New York. John wiley.
3. Wilkinson, J. H., (1965) *The Algebraic Eigenvalue Problem*. New York: Oxford Univ. Press.
4. Wilkinson, J. H. & Reinsch, C. H. (1971), *Handbook for Ahtomatic Computation. Linear Algebra, Vol*. 2, New York: Springer-Verlag.
5. Gourlay, A. R. & Watson, G. A., (1973), *Computational Methods for Matrix Eigenproblems*, New York: John Wiley.
6. Parlett, B. N., (1980). *The Symmetric Eigenvalue Problem*. Englewood Cliffs, N. J.: Prentice-Hall.
7. 曹志浩等矩阵计算和方程求根。
8. 李大潜等有限元素法续讲。
9. Moler, C. B. & Stewart, G. W (1973), "*An algorithm for generalized matrix eigenvalue Problems*" SIAM. J. Numer. Anal. 10. 241-256.
10. Stewart, G. W. (1976), "*Simultaneous iteration for computing*

invariant subspaces of non-Hermition matrices" Numer. Math. 25, 123-136.

11. Clint, M., & Jennings, A. (1971) "*A simultaneous iteration method for the unsymmetric eigenvalue Problems*" J. Inst. Maths, Applics, 8, 111-121.

第五章　线性最小二乘法

在测量平差、数据逼近以及正态分布有关的统计等问题中，都需要考虑确定一个向量$\boldsymbol{x}$，使函数

$$\rho^2(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2$$

达到最小的问题。其中A是一个$m \times n$的矩阵，$\boldsymbol{b}$是m维列向量，这个问题称为**最小二乘法问题**。本章将通过对最小二乘法的讨论来介绍$m \times n$矩阵的奇异值分解以及广义逆矩阵的某些概念和有关的递推算法。

§1　线性最小二乘法问题

在某些实际问题中碰到的线性方程组$A\boldsymbol{x} = \boldsymbol{b}$往往无解，即找不到一个向量$\boldsymbol{x}_0$，使$A\boldsymbol{x}_0 = \boldsymbol{b}$，此时只能将问题合理地改为：寻找列向量$\boldsymbol{x}$使$\rho^2(\boldsymbol{x}) = \|\boldsymbol{b} - A\boldsymbol{x}\|^2$达到最小。即使剩余向量的模的平方

$$\|\boldsymbol{r}\|^2 = \|\boldsymbol{b} - A\boldsymbol{x}\|^2 \tag{5.1.1}$$

达到最小。由于剩余向量$\boldsymbol{r}$是线性地依赖于$\boldsymbol{x}$的，所以我们称这个问题为线性最小二乘法问题。而把使(5.1.1)右端达到最小的$\boldsymbol{x}$叫做

$$A\boldsymbol{x} = \boldsymbol{b} \tag{5.1.2}$$

的最小二乘解。

现在来研究最小二乘解的存在性以及它的唯一性条件。为了以后的需要，我们首先对长方形矩阵的一些基本性质做如下讨论。

设长方形矩阵 $A\in R^{m\times n}$，$\|A\|_2$ 的定义可以仿方阵的 2 范数定义，定义为

$$\|A\|=\|A\|_2=\operatorname*{Max}_{x\neq 0}(\|Ax\|_2/\|x\|_2),\qquad(5.1.3)$$

可以证明它具有下列性质：

1) $\|A\|^2=A^TA$的最大特征值；

2) $\|A^T\|=\|A\|$；

3) $\|A^TA\|=\|A\|^2$。

对于满足方程组

$$Ax=0$$

的所有x组成的空间称为A的**零空间**，记为$N(A)$，它的维数记为$\mathrm{null}(A)$。以后也用 $R(A)$来表示A的列向量张成的线性空间。

定理5.1.1 使(5.1.1)达到最小的线性最小二乘解总是存在的，而且解唯一的充要条件是$\mathrm{null}(A)=0$

证明 根据线性代数的知识，b可唯一地分解为

$$b=b_1+b_2,$$

其中 $b_1\in R(A)$，$b_2\in R^{\perp}(A)$，$R^{\perp}(A)$是A的**正交补空间**。因为$Ax\in R(A)$，所以$-Ax\in R(A)$，故$b-Ax$有唯一分解式

$$b-Ax=b_1+b_2-Ax,\qquad(5.1.4)$$

其中$b_1-Ax\in R(A)$，$b_2\in R^{\perp}(A)$。从而可以得到

$$\|b-Ax\|^2=\|b_1-Ax\|^2+\|b_2\|^2.\qquad(5.1.5)$$

由于$b_1\in R(A)$，所以线性方程组

$$Ax=b_1\qquad(5.1.6)$$

必有解，设解为$\bar{x}$，即$b_1-A\bar{x}=0$，于是对任意向量x均有

$$\|b-Ax\|^2=\|b_1-Ax\|^2+\|b_2\|^2\geqslant\|b_2\|^2=\|b-A\bar{x}\|^2.$$

这就证明$\bar{x}$是(5.1.2)的线性最小二乘解，从而证明了线性最小二乘法问题总有解。

设 x_0 是方程组(5.1.6)的任一解，则(5.1.6)的所有解的集合可以表示为 $x_0+N(A)$。由此可得，如果 $\mathrm{null}(A)=0$，则(5.1.6)的解唯一，否则，又可得(5.1.6)的解不唯一，从而证明了(5.1.6)的解唯一存在的充要条件是 $\mathrm{null}(A)=0$。

定理5.1.2 (5.1.2)的线性最小二乘解是方程组

$$A^TAx=A^Tb \tag{5.1.7}$$

的解，反之，方程组(5.1.7)的解是(5.1.2)的最小二乘解。

证明 设 $b-Ax=b_1+b_2-Ax$，其中 $b_1\in R(A)$，$b_2\in R^\perp(A)$，若 $\bar{x}$ 是(5.1.2)的最小二乘解，则必有 $A\bar{x}=b_1$，因此，$b-A\bar{x}=b-b_1=b_2\in R^\perp(A)$。又由于 $R^\perp(A)=N(A^T)$，故 $b-A\bar{x}=b_2\in N(A^T)$，即 $A^T(b-A\bar{x})=0$，从而证明了 $\bar{x}$ 是方程组(5.1.7)的解。

反之，若 $\bar{x}$ 是方程组(5.1.7)的一个解，则必有

$$A^Tb-A^TA\bar{x}=A^T(b-A\bar{x})=0。$$

因此，

$$b-A\bar{x}\in R^\perp(A)。$$

但是，$b-A\bar{x}=b_2+b_1-A\bar{x}$，其中 $b_1-A\bar{x}\in R(A)$，$b_2\in R^\perp(A)$，根据 $b-A\bar{x}$ 在 $R(A)$、$R^\perp(A)$ 上分解的唯一性得到 $b_1-A\bar{x}=0$，于是 $\bar{x}$ 是(5.1.2)的线性最小二乘解。

定义5.1.1 方程组 $A^TAx=A^Tb$ 称为线性最小二乘法问题(5.1.1)的**法方程组**。

由定理5.1.1和定理5.1.2可知，当 $m\times n$ 矩阵 A 的零空间 $N(A)$ 的维数满足 $\mathrm{null}(A)=0$ 时，线性最小二乘解才是唯一的，即 A 的列线性无关时才有唯一的 $\bar{x}$ 使(5.1.1)达到最小。这时必有 $m\geqslant n$，而且 A^TA 是非奇异阵，因此，A^TA 是对称正定

的。所以，在null(A)=0时，求线性最小二乘解问题可以化为解系数阵为对称正定阵的法方程组(5.1.7)的问题。至于null(A)≠0的情况，线性最小二乘法问题的求解就变得复杂了。我们将分别进行讨论。

§2 法方程组的解法

本节主要讨论方程组(5.1.2)的系数阵A的列向量线性无关时(简称列满秩阵)的最小二乘解。此时，A^TA 是非奇异阵，从而方程组(5.1.7)的解可表为

$$x=(A^TA)^{-1}A^Tb.$$

若令

$$\overline{A}^+=(A^TA)^{-1}A^T, \tag{5.2.1}$$

则称矩阵$\overline{A}^+$为A的Moore-Penrose广义逆矩阵。因此，线性最小二乘解可以表为

$$x=\overline{A}^+b. \tag{5.2.2}$$

2.1 法方程组的性态

为了选择法方程组的合适的求解方法，应对法方程组(5.1.7)的特点进行研究。

形成法方程组的系数阵，需做矩阵乘积 A^TA，这不仅要花费很大的工作量，而且由于舍入误差的影响还有可能破坏 A^TA 的正定性。其次，线性方程组(5.1.7)往往是病态的。例如，给定12个点

$(-6,y_1)$，$(-5,y_2)$，…，$(-1,y_6)$，$(1,y_7)$，…，$(5,y_{11})$，$(6,y_{12})$要求确定一个8次多项式

$$y(x)=a_0+a_1x+a_2x^2+\cdots+a_8x^8$$

使$\sum(y(x_i)-y_i)^2$达到最小，这需要解方程组

$$Ca=y,$$

其中

$$C=\begin{pmatrix}1 & x_1 & x_1^2 & \cdots & x_1^8\\ 1 & x_2 & x_2^2 & \cdots & x_2^8\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 1 & x_{12} & x_{12}^2 & \cdots & x_{12}^8\end{pmatrix},$$

而 $\boldsymbol{a}=(a_0,a_1,\cdots,a_8)^T$，$\boldsymbol{y}=(y_1, y_2, \cdots,y_{12})^T$，相应的法方程组为

$$C^TC\boldsymbol{a}=C^T\boldsymbol{y}. \qquad (5.2.3)$$

设$C^TC=(c_{ij})_{9\times 9}$，则方程组(5.2.3)的条件数估计为

$$\text{Cond}(C^TC)=\frac{C^TC\text{的最大特征值}}{C^TC\text{的最小特征值}}\geqslant 4.96\times 10^{11}.$$

显然，本例中的法方程组(5.2.3)是“病态”的。因而，用一般的方法去解(5.2.3)将得不到满意的结果。

为什么法方程组(5.1.7)往往是病态的呢？现在仅就方程组(5.1.2)的右端有扰动$\delta\boldsymbol{b}$的情况来进行讨论。

设方程组(5.1.2)的右端由$\boldsymbol{b}$变为$\boldsymbol{b}+\delta\boldsymbol{b}$，此时(5.1.7)的解变为$\boldsymbol{x}+\delta\boldsymbol{x}$，且

$$A^TA(\boldsymbol{x}+\delta\boldsymbol{x})=A^T(\boldsymbol{b}+\delta\boldsymbol{b}),$$

于是有 $A^TA\delta\boldsymbol{x}=A^T\delta\boldsymbol{b}$，即

$$\delta\boldsymbol{x}=(A^TA)^{-1}A^T\delta\boldsymbol{b}=\overline{A}^+\delta\boldsymbol{b}. \qquad (5.2.4)$$

定理5.2.1 若(5.1.2)的矩阵A是列满秩阵，且$\boldsymbol{b}_1\neq 0$，则

$$\frac{\|\overline{A}^+\delta\boldsymbol{b}\|}{\|\overline{A}^+\boldsymbol{b}\|}\leqslant \text{Cond}\ (A)\ \frac{\|\delta\boldsymbol{b}_1\|}{\|\boldsymbol{b}_1\|}, \qquad (5.2.5)$$

其中$\text{Cond}(A)=\|\overline{A}^+\|\|A\|$。

证明 令$\delta\boldsymbol{b}$在R(A)和$R^\perp$(A)上的分解为

$$\delta\boldsymbol{b}=\delta\boldsymbol{b}_1+\delta\boldsymbol{b}_2,$$

其中$\delta\boldsymbol{b}_1 \in R(A)$，$\delta\boldsymbol{b}_2 \in R^T(A) = N(A^T)$，则

$$\overline{A}^+\delta\boldsymbol{b} = \overline{A}^+(\delta\boldsymbol{b}_1 + \delta\boldsymbol{b}_2) = \overline{A}^+\delta\boldsymbol{b}_1 + (A^TA)^{-1}A^T\delta\boldsymbol{b}_2 = \overline{A}^+\delta\boldsymbol{b}_1,$$

因此

$$\|\overline{A}^+\delta\boldsymbol{b}\| = \|\overline{A}^+\delta\boldsymbol{b}_1\| \leqslant \|\overline{A}^+\|\|\delta\boldsymbol{b}_1\|. \tag{5.2.6}$$

据(5.2.2)和(5.1.6)可得

$$A\overline{A}^+\boldsymbol{b} = \boldsymbol{b}_1,$$

于是

$$\|\boldsymbol{b}_1\| = A\overline{A}^+\boldsymbol{b}\| \leqslant \|A\|\|\overline{A}^+\boldsymbol{b}\|$$

$$\|\overline{A}^+\boldsymbol{b}\| \geqslant \|\boldsymbol{b}_1\|/\|A\|. \tag{5.2.7}$$

由于$\|\boldsymbol{b}_1\| \neq 0$，用(5.2.7)除(5.2.6)得到

$$\frac{\|\overline{A}^+\delta\boldsymbol{b}\|}{\|\overline{A}^+\boldsymbol{b}\|} \leqslant \|\overline{A}^+\|\ \|A\|\frac{\|\delta\boldsymbol{b}_1\|}{\|\boldsymbol{b}_1\|} = \text{Cond}\ (A)\frac{\|\delta\boldsymbol{b}_1\|}{\|\boldsymbol{b}_1\|}.$$

(5.2.2)，(5.2.4)，(5.2.5)表明，当(5.1.2)的右端有扰动$\delta\boldsymbol{b}$时，最小二乘解$\boldsymbol{x}$的相对误差界为

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leqslant \text{Cond}\ (A)\frac{\|\delta\boldsymbol{b}_1\|}{\|\boldsymbol{b}_1\|}.$$

该式的右端是$\boldsymbol{b}_1$的相对误差，因而，当(5.1.2)的右端$\boldsymbol{b}$变化时，解的变化主要依赖于$\boldsymbol{b}_1$的变化。

定理5.2.2 若阵$A \in R^{m\times n}$是列满秩阵，则

$$\text{Cond}(A^TA) = (\text{Cond}(A))^2, \tag{5.2.8}$$

其中$\text{Cond}(A^TA) = \|A^TA\|\|(A^TA)^{-1}\|$.

证明 因为$\|A\|^2 = \|A^TA\| = \|AA^T\|$，且

$$\overline{A}^+ = (A^TA)^{-1}A^T,$$

所以

$$\begin{aligned}\|\overline{A}^+\|^2 &= \|\overline{A}^+(\overline{A}^+)^T\| \\ &= \|(A^TA)^{-1}A^TA(A^TA)^{-1}\| = \|(A^TA)^{-1}\|\end{aligned}$$

故

$$\text{Cond}(A^TA)=\|A^TA\|\,\|(A^TA)^{-1}\|=\|A\|^2\|\bar{A}^+\|^2$$
$$=(\text{Cond}(A))^2.$$

定理5.2.2指出，法方程组(5.1.7)的系数矩阵的条件数是原方程组(5.1.2)的系数矩阵A的条件数的平方，即原来不太病态的问题，化为法方程组后，变得“惊人”地病态了。因此必须采用具有更高数值稳定性的方法去求解。

2.2 正交化方法

线性代数中所介绍的 Cholesky 分解是把一个对称正定阵进行三角分解，再将它应用于对A^TA的分解上，即得

$$A^TA=R^TR, \tag{5.2.9}$$

其中R是n阶非奇异的上三角阵，且规定对角元取正号。此时分解式(5.2.9)是唯一的。如果令$Q=AR^{-1}$，则

$$A=QR, \tag{5.2.10}$$

而且$Q^TQ=R^{-T}A^TAR^{-1}=R^{-T}R^TRR^{-1}=I$，即$Q$为正交阵。

对于列满秩阵A，其QR分解可由Gramm-Schmidt 正交化过程实现。并把这个正交化过程简称为G-S过程。设$A=[\boldsymbol{a}_1,\ \boldsymbol{a}_2,\ \cdots,\ \boldsymbol{a}_n]$，$Q=[\boldsymbol{q}_1,\ \boldsymbol{q}_2,\ \cdots,\ \boldsymbol{q}_n]$，则$G$-$S$过程的具体步骤如下：

1) 将$\boldsymbol{a}_1$标准化，作为Q的第1列$\boldsymbol{q}_1$，即

$$\boldsymbol{q}_1=\boldsymbol{a}_1/r_{11},\qquad r_{11}=\|\boldsymbol{a}_1\|;$$

2) 对于$k=2,3,\cdots,n$，按下列公式求出Q的相应列$\boldsymbol{q}_k$，即

$$\boldsymbol{b}_k=\boldsymbol{a}_k-\sum_{j=1}^{k-1}r_{jk}\boldsymbol{q}_j,\quad r_{jk}=(\boldsymbol{q}_j,\ \boldsymbol{a}_k),$$
$$(j=1,2,\cdots,k-1)$$
$$\boldsymbol{q}_k=\boldsymbol{b}_k/r_{kk},\qquad r_{kk}=\|\boldsymbol{b}_k\|.$$

如此求出的$\boldsymbol{q}_1,\ \boldsymbol{q}_2,\ \cdots,\ \boldsymbol{q}_n$必为标准正交向量组，这个过程可以写成如下形式：

$$
\begin{aligned}
&\boldsymbol{a}_1 = r_{11}\boldsymbol{q}_1,\\
&\boldsymbol{a}_2 = r_{12}\boldsymbol{q}_1 + r_{22}\boldsymbol{q}_2,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\boldsymbol{a}_n = r_{1n}\boldsymbol{q}_1 + r_{2n}\boldsymbol{q}_2 + \cdots + r_{nn}\boldsymbol{q}_n.
\end{aligned}
\tag{5.2.11}
$$

将(5.2.11)写成矩阵形式，即为$A=QR$，其中

$$
R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{nn} \end{pmatrix},
$$

Q为正交阵.

将分解式(5.2.10)代入方程组(5.1.7)，得到

$$R^TQ^TQR\boldsymbol{x} = R^TQ^T\boldsymbol{b}.$$

由于R^T是非奇异阵，所以存在逆阵$(R^T)^{-1}$，用它左乘于上式两端，得到

$$R\boldsymbol{x} = Q^T\boldsymbol{b}. \tag{5.2.12}$$

因此，求得R和Q后，用$Q^T\boldsymbol{b}$作为三角形方程组(5.2.12)的右端项，然后解出此三角形方程组，即可求得最小二乘解$\boldsymbol{x}$.

实际使用G-S过程时，由于Q的列是由A的列的线性组合产生的，一般说来，舍入误差较大，以致严重地影响Q阵的正交性，从而该算法的数值稳定性可能很差。为了提高算法的精度，将G-S过程进行适当修改，便得到所谓“修改的G-S正交化过程”.

算法5.2.1　修改的Gramm-Schmidt方法(MGS方法).

具体步骤如下.

1)　对$k=1,2,\cdots,n-1$，做

1.1)　$\boldsymbol{q}_k = \boldsymbol{a}_k/r_{kk}$,　　$r_{kk}=\|\boldsymbol{a}_k\|$,　　(5.2.13)

1.2)　对$j=k+1,\cdots,n$做

1.2.1) $a_j \leftarrow a_j - r_{kj} q_k,\ r_{kj} = (q_k,\ a_j)$ (5.2.14)

1.2.2) NEXT j,

1.3) NEXT k.

2) 置$Q=(q_1,\cdots,q_n)$, $R=(r_1,\cdots,r_n)$, (5.2.15)

$$r_j^T=(r_{1j},r_{2j},\cdots,r_{jj},0,\cdots,0).$$

3) 解方程组 $Rx=Q^Tb$. (5.2.16)

4) 输出 $x=R^{-1}Q^Tb$. (5.2.17)

2.3 镜像变换法

我们知道，正交变换下任何向量的欧氏模是不变的。因此，如果Q是正交阵，那么

$$\|b-Ax\|=\|Q(b-Ax)\|=\|Qb-QAx\|,$$

于是求x使$\|b-Ax\|$达到最小的问题与求x使$\|Qb-QAx\|$达到最小的问题是等价的，其解均为最小二乘解。

根据第一章§3的理论，当A是列满秩矩阵时，存在正交阵Q，使

$$QA=\begin{pmatrix} R \\ \\ 0 \end{pmatrix}, \tag{5.2.18}$$

其中R是$n\times n$的上三角阵。令

$$Qb=\begin{pmatrix} b_1 \\ \\ b_2 \end{pmatrix}\begin{matrix} \}n \\ \\ \}m-n, \end{matrix}$$

于是

$$\|b-Ax\|=\left\|\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}-\begin{bmatrix} R \\ 0 \end{bmatrix}x\right\|$$

$$=[(b_1-Rx)^T(b_1-Rx)+b_2^Tb_2]^{\frac{1}{2}}. \tag{5.2.19}$$

显然，$b_1-Rx=0$的解便是使(5.2.19)达到最小的最小二乘

解。因此，不需要解法方程组(5.1.7)，而只需要解一次三角形方程组 $Rx=b$，便可求出线性最小二乘解。

在 A 为列满秩阵的情况下，使用正交化方法或镜像变换法解“矛盾不太大的方程组”时，由于

$$\begin{aligned}\text{Cond}(R) &= \sqrt{\frac{R^TR\text{ 的最大特征值}}{R^TR\text{ 的最小特征值}}} \\ &= \sqrt{\frac{R^TQ^TQR\text{ 的最大特征值}}{R^TQ^TQR\text{ 的最小特征值}}} \\ &= \text{Cond}(QR) = \text{Cond}(A),\end{aligned}$$

即保持了条件数不变，从而避免了处理更为病态的方程组，因此用以上的方法去解最小二乘法问题，一般来说是可以获得比较满意的结果的。使用正交化方法和镜像变换法解矛盾方程组的具体算法，可以参看 Stewart 著的“矩阵计算引论”第五章。

§3 奇异值分解及广义逆矩阵

前面我们研究了在 $\text{null}(A)=0$ 时求 $Ax=b$ 的最小二乘解问题。当 $\text{null}(A)\neq 0$ 即矩阵 A 的秩 $r<n$ 时，$Ax=b$ 的最小二乘解无论在解的表示形式和求解的方法上都将更为复杂，为了对后者进行较详细的研究，本节将引进**奇异值分解**和**广义逆矩阵**等概念。

3.1 奇异值分解

引理5.3.1 设矩阵 $A\in R^{m\times n}$ 的秩为 r，则存在正交矩阵 $U\in R^{m\times m}$，$V\in R^{n\times n}$，使

$$U^TAV=\begin{bmatrix}\Delta & 0\\ 0 & 0\end{bmatrix}, \tag{5.3.1}$$

其中 Δ 是非奇异的 r 阶下三角阵。

证明 对于秩为 r 的矩阵 $A\in R^{m\times n}$，存在正交阵 Q 和排列

阵 P；它们分别属于 $R^{m\times m}$ 和 $R^{n\times n}$ 使

$$QAP=\begin{bmatrix}\overline{A}_{11} & \overline{A}_{12}\\ 0 & 0\end{bmatrix},$$

其中 $\overline{A}_{11}$ 是 r 阶非奇异的上三角阵，$\overline{A}_{12}$ 是 $r\times(n-r)$ 矩阵。有

$$(QAP)^T=\begin{bmatrix}\overline{A}_{11}^T & 0\\ \overline{A}_{12}^T & 0\end{bmatrix}.$$

对此又存在正交阵 $\overline{Q}\in R^{n\times n}$，使

$$\overline{Q}(QAP)^T=\begin{bmatrix}\Delta^T & 0\\ 0 & 0\end{bmatrix},$$

其中 Δ^T 是 r 阶非奇异的上三角阵，因此

$$(\overline{Q}^T(QAP)^T)^T=\begin{bmatrix}\Delta & 0\\ 0 & 0\end{bmatrix},$$

即

$$QAP\overline{Q}^T=\begin{bmatrix}\Delta & 0\\ 0 & 0\end{bmatrix}.$$

令 $Q^T=U$，$P\overline{Q}^T=V$。因为 Q，$\overline{Q}$，P 均为正交阵，所以 U，V 也是正交阵，且分别属于 $R^{m\times m}$ 和 $R^{n\times n}$，使

$$U^TAV=\begin{bmatrix}\Delta & 0\\ 0 & 0\end{bmatrix},$$

其中 Δ 是 r 阶非奇异的下三角阵。

定理5.3.1 设矩阵 $A\in R^{m\times n}$ 的秩为 r，则存在 m 阶正交阵 U 和 n 阶正交阵 V，使

$$U^TAV=D,\quad A=UDV^T,\tag{5.3.2}$$

$$D=\begin{bmatrix}D_r & 0\\ 0 & 0\end{bmatrix},\quad D_r=\mathrm{diag}(\mu_1,\mu_2,\cdots,\mu_r),$$

其中 $\mu_i=\sqrt{\lambda_i}$，而 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r>0$ 是 A^TA 的非零特征值的全体。

证明 由引理5.3.1知道，存在正交阵$\overline{U}$和$\overline{V}$使

$$\overline{U}^T A \overline{V} = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix},$$

其中Δ是r阶非奇异下三角阵。因此$\Delta^T\Delta$是r阶对称正定阵，从而存在正交阵V_r使

$$\Delta^T \Delta = V_r \Lambda_r V_r^T$$

$$\Lambda_r = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r \end{pmatrix},$$

这里$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$是$\Delta^T\Delta$的全部特征值。它们也是$A^TA$的非零特征值。令

$$D_r = \begin{pmatrix} \mu_1 & & & \\ & \mu_2 & & \\ & & \ddots & \\ & & & \mu_r \end{pmatrix}, \quad \mu_i = \sqrt{\lambda_i} \ (i = 1, 2, \cdots, r)$$

则有 $D_r^2 = \Lambda_r$，设

$$U_r = \Delta V_r D_r^{-1}, \tag{5.3.3}$$

则

$$U_r^T U_r = (D_r^{-1})^T V_r^T \Delta^T \Delta V_r D_r^{-1}$$
$$= D_r^{-1} V_r^T V_r \Lambda_r V_r^T V_r D_r^{-1} = D_r^{-1} \Lambda_r D_r^{-1} = I_r$$

即U_r是正交阵，从而

$$\overline{\overline{U}} = \begin{bmatrix} U_r & 0 \\ 0 & I_{m-r} \end{bmatrix}, \quad \overline{\overline{V}} = \begin{bmatrix} V_r & 0 \\ 0 & I_{n-r} \end{bmatrix}$$

分别为m和n阶正交阵。又由(5.3.3)可得

$$U_r^T \Delta V_r = D_r$$

及

$$\overline{\overline{U}}^T \overline{U}^T A \overline{V}\, \overline{\overline{V}} = \overline{\overline{U}}^T \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \overline{\overline{V}}$$

$$=\begin{bmatrix} U_r^T\Delta V_r & 0 \\ 0 & 0 \end{bmatrix}=\begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}. \tag{5.3.4}$$

令$\bar{U}\bar{\bar{U}}=U$，$\bar{V}\bar{\bar{V}}=V$，则U，V都是正交阵，且

$$U^TAV=\begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix},$$

其中$D_r=\mathrm{diag}(\mu_1,\mu_2,\cdots,\mu_r)$。

定义5.3.1 设矩阵$A\in R^{m\times n}$的秩为r，A^TA的非零特征值为$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r>0$，则$\mu_i=\sqrt{\lambda_i}\ (i=1,2,\cdots,r)$称为$A$的奇异值，等式$A=UDV^T$称为$A$的奇异值分解。

在A的奇异值分解$A=UDV^T$中，D是由A的奇异值唯一确定的，但是正交阵U和V的选取却不是唯一的。当$\Delta^T\Delta$有重特征值时，r阶正交阵V_r的选取就不唯一，从而U和V的选取也就不唯一。当矩阵A是对称阵时，A的奇异值就是它的特征值的绝对值。

有了引理5.3.1和定理5.3.1以后，我们就可以在一般的情况下来讨论最小二乘法问题的解法了。由于

$$\|\boldsymbol{b}-A\boldsymbol{x}\|=\|U^T\boldsymbol{b}-U^TA\boldsymbol{x}\|=\|U^T\boldsymbol{b}-U^TAVV^T\boldsymbol{x}\|$$

$$=\left\|U^T\boldsymbol{b}-\begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}V^T\boldsymbol{x}\right\|, \tag{5.3.5}$$

令 $$\boldsymbol{c}=U^T\boldsymbol{b}=\begin{bmatrix} \boldsymbol{c}_r \\ \boldsymbol{c}_{n-r} \end{bmatrix},\quad \boldsymbol{y}=V^T\boldsymbol{x}=\begin{bmatrix} \boldsymbol{y}_r \\ \boldsymbol{y}_{n-r} \end{bmatrix}, \tag{5.3.6}$$

将(5.3.6)代入(5.3.5)后得到

$$\|\boldsymbol{b}-A\boldsymbol{x}\|^2=\left\|\begin{bmatrix} \boldsymbol{c}_r \\ \boldsymbol{c}_{n-r} \end{bmatrix}-\begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{y}_r \\ \boldsymbol{y}_{n-r} \end{bmatrix}\right\|^2$$

$$=\left\|\begin{bmatrix} \boldsymbol{c}_r-\Delta\boldsymbol{y}_r \\ \boldsymbol{c}_{n-r} \end{bmatrix}\right\|^2=\|\boldsymbol{c}_r-\Delta\boldsymbol{y}_r\|^2+\|\boldsymbol{c}_{n-r}\|^2.$$

因为Δ是非奇异的下三角阵，所以方程组

$$\boldsymbol{c}_r - \Delta \boldsymbol{y}_r = 0 \tag{5.3.7}$$

有唯一解 $\bar{\boldsymbol{y}}_r$，令

$$\bar{\boldsymbol{x}} = V\begin{pmatrix} \bar{\boldsymbol{y}}_r \\ \bar{\boldsymbol{y}}_{n-r} \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{x}}_r \\ \bar{\boldsymbol{x}}_{n-r} \end{pmatrix}, \tag{5.3.8}$$

其中 $\bar{\boldsymbol{y}}_{n-r}$ 是任意 $(n-r)$ 维列向量．由于

$$\|\boldsymbol{b} - A\boldsymbol{x}\|^2 \geqslant \|\boldsymbol{c}_{n-r}\|^2 = \|\boldsymbol{c}_r - \Delta\bar{\boldsymbol{y}}_r\|^2 + \|\boldsymbol{c}_{n-r}\|^2$$
$$= \|\boldsymbol{b} - A\bar{\boldsymbol{x}}\|^2,$$

故 $\bar{\boldsymbol{x}}$ 是(5.1.2)的最小二乘解．

如果取 $\bar{\boldsymbol{y}}_{n-r} = 0$，令

$$\bar{\boldsymbol{x}}^{(0)} = V\begin{pmatrix} \bar{\boldsymbol{y}}_r \\ 0 \end{pmatrix}, \tag{5.3.9}$$

则

$$\|\bar{\boldsymbol{x}}\|^2 = \left\|\begin{pmatrix} \bar{\boldsymbol{x}}_r \\ \bar{\boldsymbol{x}}_{n-r} \end{pmatrix}\right\|^2 = \|\bar{\boldsymbol{x}}_r\|^2 + \|\bar{\boldsymbol{x}}_{n-r}\|^2 \geqslant \|\bar{\boldsymbol{x}}_r\|^2 = \|\bar{\boldsymbol{x}}^{(0)}\|^2,$$

所以(5.3.9)是最小二乘解中向量模为最小的解．

如果用 $A = UDV^T$ 代入(5.3.5)中，就得到

$$\|\boldsymbol{b} - A\boldsymbol{x}\| = \|U^T\boldsymbol{b} - U^TUDV^T\boldsymbol{x}\| = \|U^T\boldsymbol{b} - DV^T\boldsymbol{x}\|. \tag{5.3.10}$$

再将(5.3.6)代入(5.3.10)，得到

$$\|\boldsymbol{b} - A\boldsymbol{x}\|^2 = \left\|\begin{bmatrix} \boldsymbol{c}_r \\ \boldsymbol{c}_{n-r} \end{bmatrix} - \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{y}_r \\ \boldsymbol{y}_{n-r} \end{bmatrix}\right\|^2$$
$$= \|\boldsymbol{c}_r - D_r\boldsymbol{y}_r\|^2 + \|\boldsymbol{y}_{n-r}\|^2.$$

类似于前面的讨论，可以得到(5.1.2)的向量模最小的最小二乘解

$$\bar{\boldsymbol{x}}^{(0)} = V\begin{bmatrix} D_r^{-1}\boldsymbol{c}_r \\ 0 \end{bmatrix} = V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{c}_r \\ \boldsymbol{c}_{n-r} \end{bmatrix}$$
$$= V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^T\boldsymbol{b}. \tag{5.3.11}$$

从这个表达式中，我们可以看到(5.1.2)的最小二乘解等于矩阵

$$V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^T \tag{5.3.12}$$

与列向量 $\boldsymbol{b}$ 的乘积。而矩阵(5.3.12)与第二节中引入 $\overline{A}^+$ 有同样的作用，我们也称之为 $\boldsymbol{A}$ 的广义逆矩阵。

3.2 广义逆矩阵

定义 5.3.2 对于矩阵 $A \in R^{m\times n}$，当存在满足如下四个条件（称 Penrose 条件）的 $n\times m$ 矩阵 X 时，称 X 为 A 的广义逆矩阵，记为 A^+

$$AXA = A, \qquad (p_1)$$

$$XAX = X, \qquad (p_2)$$

$$AX = (AX)^T, \qquad (p_3)$$

$$XA = (XA)^T. \qquad (p_4)$$

根据广义逆矩阵的定义，显然它具有如下性质：

1) $(A^+A)^2 = A^+A$;

2) $(AA^+)^2 = AA^+$.

定理5.3.2 设矩阵 $A \in R^{m\times n}$ 的奇异值分解为

$$A = UDV^T, \qquad D = \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix},$$

则

$$X = V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^T$$

为 A 的广义逆矩阵 A^+，而且是唯一的。

证明 因为

$$AXA = AV\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^TA = UDV^TV\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^TUDV^T$$

$$= UDV^T = A,$$

$$XAX = V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T U D V^T V \begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T$$

$$= V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T = X,$$

故

$$AX = UDV^T V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T = U\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^T = (AX)^T.$$

类似可证　$XA = (XA)^T$.

所以X是A的广义逆矩阵A^+，即

$$A^+ = V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T. \tag{5.3.13}$$

如果X和$\overline{X}$均为满足$(p_1) \sim (p_4)$的广义逆矩阵，则

$$X = XAX = XA\overline{X}AX = X(AXA\overline{X})^T$$

$$= X(A\overline{X})^T = XA\overline{X} = XA\overline{X}A\overline{X} = (XA)^T(\overline{X}A)^T\overline{X}$$

$$= (\overline{X}AXA)^T = XAX = X.$$

从而证明了A^+的存在和唯一性.

从广义逆矩阵A^+的唯一性知

$$A^+ = V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^T,$$

它将不依赖于A的奇异值分解中的正交阵U和V的选取.

推理1　若矩阵$A \in R^{m\times n}$的广义逆矩阵为A^+，则A^T的广义逆矩阵为$(A^+)^T$，即

$$(A^+)^T = (A^T)^+.$$

证明　设$A = UDV^T$,　　$D = \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}$,

则

$$A^T = VDU^T = V\begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} U^T,$$

从而

$$(A^T)^+ = (U^T)^T\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}V^T = \left(V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^T\right)^T$$

$$= (A^+)^T,$$

即

$$(A^T)^+ = (A^+)^T.$$

推理2 $(A^+)^+ = A$。

其证明作为习题。

当A为列满秩阵时，可以证明§2中定义的广义逆矩阵$\overline{A}^+ = (A^TA)^{-1}A^T$与本节定义的广义逆矩阵是一样的，即$\overline{A}^+ = A^+$。

定理5.3.3 若A是列满秩阵，则

$$(A^TA)^{-1}A^T = A^+;$$

若A是行满秩阵，则

$$A^T(AA^T)^{-1} = A^+.$$

根据定义5.3.2和定理5.3.2，只要证明$(A^TA)^{-1}A^T$满足Penrose条件就行了。具体证明作为练习。

推理1 当A为列满秩阵时，有$A^+A = I$；当A为行满秩阵时，有$AA^+ = I$。

推理2 当n阶方阵A为满秩时，则$A^+ = A^{-1}$。

这个推论说明了A^+是A^{-1}的推广。

定理5.3.3指出，当$A\in R^{m\times n}$的秩为n或m时，广义逆矩阵可以表示成另一种形式。现在我们将讨论当秩为r时，如何用不同于(5.3.13)的形式来表示A的广义逆矩阵的问题。先证明如下定理。

定理5.3.4 对于秩为$r\neq 0$的矩阵$A\in R^{m\times n}$，必有列满秩阵$B\in R^{m\times r}$与行满秩阵$C\in R^{r\times n}$，使

$$A = BC.$$

证明 对于秩为 $r \neq 0$ 的矩阵 A，必有满秩矩阵 P 和 Q 使

$$PAQ = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \tag{5.3.14}$$

其中 I_r 是 r 阶单位矩阵。设

$$P^{-1} = [B\bar{B}], \qquad Q^{-1} = \begin{bmatrix} C \\ \bar{C} \end{bmatrix},$$

于是有

$$\begin{aligned} A &= P^{-1}(PAQ)Q^{-1} \\ &= [B \quad \bar{B}] \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} C \\ \bar{C} \end{bmatrix} = BC. \end{aligned}$$

因为 P^{-1} 的 m 个列线性无关，而 B 是由 P^{-1} 中的 r 个列组成，所以 B 是列满秩阵，同理 C 是行满秩阵。

定义5.3.3 对秩为 $r \neq 0$ 的矩阵 $A \in R^{m \times n}$，称分解式

$$A = BC \tag{5.3.15}$$

为 A 的**秩分解式**，其中 B 和 C 分别是列满秩阵和行满秩阵。

定义5.3.4 若 $A = BC$ 是非零矩阵 A 的秩分解式，则

$$C^T(CC^T)^{-1}(B^TB)^{-1}B^T \tag{5.3.16}$$

叫做 A 的一个**广义逆矩阵**。若 $A = O$，则 O 就是 A 的广义逆矩阵。按定理5.3.3的记法，(5.3.16)可以写成

$$C^+B^+. \tag{5.3.17}$$

定理5.3.5 A 的广义逆矩阵 C^+B^+ 满足 Penrose 条件 $(P_1) \sim (P_4)$。

证明 若记 $X = C^+B^+$ 则

$$AXA = BCC^T(CC^T)^{-1}(B^TB)^{-1}B^TBC = BC = A,$$

$$\begin{aligned} AX &= BCC^T(CC^T)^{-1}(B^TB)^{-1}B = B(B^TB)^{-1}B^T \\ &= (B(B^TB)^{-1}B^T)^T = (AX)^T. \end{aligned}$$

同理可证 $XAX = X$，$XA = (XA)^T$。

对于矩阵 A，因为满足 Penrose 条件 $(P_1)\sim(P_4)$ 的广义逆矩阵是唯一的，所以广义逆矩阵 C^+B^+ 可记为

$$A^+=C^+B^+,$$

即

$$C^+B^+=V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^T.$$

例5.3.1 求

$$\begin{cases} x-y=-1, \\ x-y=1, \\ 2x-y=0 \end{cases}$$

的最小二乘解.

解 设

$$B=\begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 2 & -1 \end{pmatrix},$$

显然 B 的秩为 2，所以

$$B^+=(B^TB)^{-1}B^T=\frac{1}{2}\begin{bmatrix} -1 & -1 & 2 \\ -2 & -2 & 2 \end{bmatrix},$$

于是

$$\begin{pmatrix} x \\ y \end{pmatrix}=B^+\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}=\frac{1}{2}\begin{pmatrix} -1 & -1 & 2 \\ -2 & -2 & 2 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

例5.3.2 求 $Ax=b$ 的模为最小的最小二乘解，其中

$$A=\begin{pmatrix} -1 & 2 & 1 \\ -1 & 2 & 1 \\ 0 & 3 & 2 \end{pmatrix},\qquad b=\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

解 显然 A 的秩为2，它的一个秩分解为

$$A=\begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 2 & -1 \end{pmatrix}\cdot\begin{bmatrix} 1 & 1 & 1 \\ 2 & -1 & 0 \end{bmatrix},$$

$$B^{+} = (B^{T}B)^{-1}B^{T} = \frac{1}{2}\begin{bmatrix} -1 & -1 & 2 \\ -2 & -2 & 2 \end{bmatrix},$$

$$C^{+} = C^{T}(CC^{T})^{-1} = \frac{1}{14}\begin{pmatrix} 3 & 5 \\ 6 & -4 \\ 5 & -1 \end{pmatrix}$$

$$A^{+} = C^{+}B^{+} = \frac{1}{14}\begin{pmatrix} 3 & 5 \\ 6 & -4 \\ 5 & -1 \end{pmatrix}\frac{1}{2}\begin{bmatrix} -1 & -1 & 2 \\ -2 & -2 & 2 \end{bmatrix}$$

$$= \frac{1}{28}\begin{pmatrix} -13 & -13 & 16 \\ 2 & 2 & 4 \\ -3 & -3 & 8 \end{pmatrix},$$

于是所求的解为

$$\boldsymbol{x} = A^{+}\boldsymbol{b} = \frac{1}{28}\begin{pmatrix} -13 & -13 & 16 \\ 2 & 2 & 4 \\ -3 & -3 & 8 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{28}\begin{pmatrix} 16 \\ 4 \\ 8 \end{pmatrix} = \frac{1}{7}\begin{pmatrix} 4 \\ 1 \\ 2 \end{pmatrix}.$$

下面我们用广义逆矩阵来给出方程组$A\boldsymbol{x}=\boldsymbol{b}$的解及线性最小二乘解的一般表达式。

定理5.3.6 线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 有解的充要条件是

$$AA^{+}\boldsymbol{b} = \boldsymbol{b} \tag{5.3.18}$$

当方程组有解时，它的通解是

$$\boldsymbol{x} = A^{+}\boldsymbol{b} + (I - A^{+}A)\boldsymbol{\xi}, \tag{5.3.19}$$

其中$\boldsymbol{\xi}$是任意n维列向量。

当方程组无解时，(5.3.19)是最小二乘法的通解（可理解为对应法方程组的通解），取$\xi=0$，即得模为最小的最小二乘解$A^{+}\boldsymbol{b}$。

证明 1) 若方程组有解，即存在$\boldsymbol{y}$使$A\boldsymbol{y}=\boldsymbol{b}$，由于$A=AA^{+}A$，故

$$b = Ay = AA^+Ay = AA^+b,$$

即 (5.3.18) 成立。反之，若 $AA^+b = b$，则 $A^+b = x$ 便是一个解，从而证明了方程组有解的充要条件是$AA^+b = b$。

2) 当方程组有解时，显然 $x^{(0)} = A^+b$ 是它的一个解。现在要证齐次方程组$Ax = 0$的通解是

$$\bar{x} = (I - A^+A)\xi$$

其中 ξ 是任意 n 维列向量。

因为 $A\bar{x} = A(I - A^+A)\xi = (A - AA^+A)\xi = 0,$

故 $\bar{x} = (I - A^+A)\xi$是 $Ax = 0$ 的解，另一方面，$Ax = 0$的任何解 x 可表为

$$x = x - A^+Ax = (I - A^+A)x.$$

又满足上式的 x 必然满足$Ax = 0$，从而证明了$Ax = 0$ 的通解是

$$\bar{x} = (I - A^+A)\xi$$

其中 ξ 是任意 n 维列向量，综上所述，$Ax = b$ 的通解为

$$x = A^+b + (I - A^+A)\xi.$$

3) 根据定理 5.1.1 和定理 5.1.2可知，最小二乘法必有解，且其解必为法方程组

$$A^TAx = A^Tb$$

的解。由 2) 知道，该方程组的通解是

$$x = (A^TA)^+A^Tb + [I - (A^TA)^+(A^TA)]\xi,$$

其中 ξ 是任意 n 维列向量。

设A 的奇异值分解为$A = UDV^T$，则

$$A^TA = VDU^TUDV^T = VD^2V^T$$

是A^TA的奇异值分解，所以

$$(A^TA)^+ = V\begin{bmatrix} (D_r^{-1})^2 & 0 \\ 0 & 0 \end{bmatrix}V^T,$$

于是

$$(A^TA)^+A^T = V\begin{bmatrix} (D_r^{-1})^2 & 0 \\ 0 & 0 \end{bmatrix}V^TVDU^T$$

$$= V\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix}U^T = A^+,$$

$$(A^TA)^+(A^TA) = ((A^TA)^+A^T)A = A^+A,$$

故法方程组的通解是

$$x = A^+b + (I - A^+A)\xi,$$

其中ξ是任意n维向量。当$\xi = 0$时，$x = A^+b$，由广义逆矩阵的唯一性和（5.3.11）可知，$x = A^+b$就是向量模为最小的最小二乘解。

§4 广义逆矩阵的一个递推算法

4.1 对称分块矩阵A的求逆

设对称分块矩阵为

$$A = \begin{bmatrix} P & R^T \\ R & Q \end{bmatrix},$$

其中P、Q也为对称矩阵且逆矩阵P^{-1}、A^{-1}存在。如果

$$\begin{bmatrix} P & R^T \\ R & Q \end{bmatrix}\begin{bmatrix} x_p \\ x_q \end{bmatrix} = \begin{bmatrix} b_p \\ b_q \end{bmatrix}, \tag{5.4.1}$$

则

$$\begin{bmatrix} x_p \\ x_q \end{bmatrix} = \begin{bmatrix} P & R^T \\ R & Q \end{bmatrix}^{-1}\begin{bmatrix} b_p \\ b_q \end{bmatrix}. \tag{5.4.2}$$

将(5.4.1)写成矩阵方程组的形式，有

$$\begin{cases}P\boldsymbol{x}_p+R^T\boldsymbol{x}_q=\boldsymbol{b}_p, & (5.4.3)\\ R\boldsymbol{x}_p+Q\boldsymbol{x}_q=\boldsymbol{b}_q. & (5.4.4)\end{cases}$$

类似于解二元一次方程组那样，由(5.4.3)可得

$$\boldsymbol{x}_p=-P^{-1}(R^T\boldsymbol{x}_q-\boldsymbol{b}_p), \tag{5.4.5}$$

将(5.4.5)代入(5.4.4)，化简后可得

$$(Q-RP^{-1}R^T)\boldsymbol{x}_q=\boldsymbol{b}_q-RP^{-1}\boldsymbol{b}_p,$$

由于它的解存在且唯一，所以

$$\boldsymbol{x}_q=-(Q-RP^{-1}R^T)^{-1}RP^{-1}\boldsymbol{b}_p+(Q-RP^{-1}R^T)^{-1}\boldsymbol{b}_q. \tag{5.4.6}$$

令 $Z=(Q-RP^{-1}R^T)^{-1}$，$Y=-ZRP^{-1}$，代入(5.4.6)，得到

$$\boldsymbol{x}_q=Y\boldsymbol{b}_p+Z\boldsymbol{b}_q. \tag{5.4.7}$$

将(5.4.7)代入(5.4.5)化简后，得到

$$\boldsymbol{x}_p=(P^{-1}-P^{-1}R^TY)\boldsymbol{b}_P+Y^T\boldsymbol{b}_q. \tag{5.4.8}$$

令 $W=P^{-1}-P^{-1}R^TY$，代入(5.4.8)得到

$$\boldsymbol{x}_p=W\boldsymbol{b}_p+Y^T\boldsymbol{b}_q. \tag{5.4.9}$$

将(5.4.9)和(5.4.7)写成矩阵形式，即得所求的解

$$\begin{bmatrix}\boldsymbol{x}_p\\ \boldsymbol{x}_q\end{bmatrix}=\begin{bmatrix}W & Y^T\\ Y & Z\end{bmatrix}\begin{bmatrix}\boldsymbol{b}_p\\ \boldsymbol{b}_q\end{bmatrix}. \tag{5.4.10}$$

比较(5.4.10)和(5.4.2)得到

$$\begin{bmatrix}P & R^T\\ R & Q\end{bmatrix}^{-1}=\begin{bmatrix}W & y^T\\ Y & Z\end{bmatrix}, \tag{5.4.11}$$

其中

$$\begin{cases}Z=(Q-RP^{-1}R^T)^{-1},\\ Y=-ZRP^{-1},\\ W=P^{-1}-P^{-1}R^TY.\end{cases} \tag{5.4.12}$$

从 (5.4.12) 我们可以看出：分块求逆的方法是：先求出 P^{-1}，再对 $(Q-RP^{-1}R^T)$ 求逆，然后利用矩阵的乘法求得 y 和 W，由此便可得到 A 的逆矩阵 A^{-1}。

例5.4.1 设

$$A=\begin{bmatrix} P & R^T \\ R & Q \end{bmatrix},$$

其中

$$P=Q=\begin{pmatrix} 3 & 0 & -1 & 0 \\ 0 & 3 & 0 & -1 \\ -1 & 0 & 3 & -1 \\ 0 & -1 & -1 & 3 \end{pmatrix},\quad R=\begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

求 A 的逆矩阵 A^{-1}。

解 设

$$A^{-1}=\begin{bmatrix} W & Y^T \\ Y & Z \end{bmatrix},$$

由已知条件先求得

$$P^{-1}=\frac{1}{55}\begin{pmatrix} 21 & 1 & 8 & 3 \\ 1 & 21 & 3 & 8 \\ 8 & 3 & 24 & 9 \\ 3 & 8 & 9 & 24 \end{pmatrix},\quad RP^{-1}=-\frac{1}{55}\begin{pmatrix} 8 & 3 & 24 & 9 \\ 3 & 8 & 9 & 24 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$RP^{-1}R^T=\frac{3}{55}\begin{pmatrix} 8 & 3 & 0 & 0 \\ 3 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

故

$$Z=(Q-RP^{-1}R^T)^{-1}=\frac{1}{2071}\begin{pmatrix} 963 & 127 & 377 & 168 \\ 127 & 963 & 168 & 377 \\ 377 & 168 & 939 & 369 \\ 168 & 377 & 369 & 939 \end{pmatrix};$$

$$Y = -ZRP^{-1} = \frac{1}{2071}\begin{pmatrix} 147 & 74 & 441 & 213 \\ 71 & 147 & 213 & 441 \\ 64 & 45 & 192 & 135 \\ 45 & 64 & 135 & 192 \end{pmatrix},$$

由 Y 便可求得 Y^T；

$$W = P^{-1} - P^{-1}R^TY = \frac{1}{2071}\begin{pmatrix} 816 & 56 & 377 & 168 \\ 56 & 816 & 168 & 377 \\ 377 & 168 & 1131 & 504 \\ 168 & 377 & 504 & 1131 \end{pmatrix}.$$

由以上几项，A^{-1} 便可得知。

4.2 计算广义逆矩阵的递推算法

引理5.4.1 设

$$B = \begin{bmatrix} I_r \\ pC^+ \end{bmatrix},$$

其中 p 是 n 维行向量，C^+是 $n\times r$ 阵，则

$$(BC)^+ = C^+B^+.$$

证明 设$BC = A$，$C^+B^+ = X$，由假设可得

$$B^TB = I_r + (PC^+)^TPC^+. \qquad (5.4.13)$$

因为 B 是列满秩阵，所以$B^+B = I$。因此，

$$\begin{aligned} XA &= C^+B^+(BC) = C^+B^+(BC)(BC)^+BC = C^+C(BC)^+BC \\ &= C^T(C^+)^TC^TB^T((BC)^+)^T = C^TB^T((BC)^+)^T \\ &= [(BC)^T((BC)^+)^T]^T = (BC)^+(BC) = A^+A, \qquad (5.4.14) \end{aligned}$$

于是

$$(XA)^T = (A^+A)^T = A^+A = XA, \quad AXA = AA^+A = A.$$

使用(5.4.13)，可以类似地证明

$$AX = AA^+, \qquad (5.4.15)$$

于是

$$(AX)^T=(AA^+)^T=AA^+=AX,$$

$$XAX=C^+B^+BCC^+B^+=C^+CC^+B^+=C^+B^+=X.$$

因此，$X=C^+B^+$ 满足 Penrose 条件，据定义 5.3.2，X是A的广义逆矩阵，即$(BC)^+=C^+B^+$。

引理5.4.2 设

$$A=\begin{bmatrix}D & O\\ O & I\end{bmatrix}\begin{bmatrix}Q\\ p\end{bmatrix},$$

其中D，Q分别属于$R^{k\times t}$和$R^{t\times n}(t<n)$且$DQ=C$为C的秩分解，P为n维行向量并与Q的t行线性无关，且$d=p(I-C^+C)\neq 0$，则

$$A^+=(C^+-d^+pC^+d^+).\qquad(5.4.16)$$

证明 因为

$$A=\begin{bmatrix}D & O\\ O & 1\end{bmatrix}\begin{bmatrix}Q\\ p\end{bmatrix}$$

是A的一个秩分解，且

$$\begin{bmatrix}D & O\\ O & 1\end{bmatrix}^+=\begin{bmatrix}D^+ & O\\ O & 1\end{bmatrix},$$

所以只需求

$$\begin{bmatrix}Q\\ p\end{bmatrix}^+.$$

据定理5.3.3

$$\begin{bmatrix}Q\\ p\end{bmatrix}^+=[Q^T,p^T]\begin{bmatrix}QQ^T & Qp^T\\ pQ^T & pp^T\end{bmatrix}^{-1}=[Q^T,p^T]\begin{bmatrix}W & Y\\ Y^T & Z\end{bmatrix}$$

$$=[Q^TW+p^Ty^T,\ Q^Ty+p^TZ].\qquad(5.4.17)$$

由对称分块矩阵求逆法知道：

$$Z=[pp^T-pQ^T(QQ^T)^{-1}Qp^T]^{-1}=[pp^T-pQ^+Qp^T]^{-1},$$

$$Y^T=-ZpQ^T(QQ^T)^{-1}=-ZpQ^+,$$

$$W=(QQ^T)^{-1}-(QQ^T)^{-1}Qp^TY^T,$$

因为
$$C^+C = Q^+D^+DQ = Q^+Q,$$
而且
$$(C^+C)^T = C^+C,$$
$$\begin{aligned}\mathbf{d}\mathbf{d}^T &= \mathbf{p}(I - C^+C)(I - C^+C)^T\mathbf{p}^T = \mathbf{p}(I - C^+C)\mathbf{p}^T \\ &= \mathbf{p}\mathbf{p}^T - \mathbf{p}C^+C\mathbf{p}^T = \mathbf{p}\mathbf{p}^T - \mathbf{p}Q^+Q\mathbf{p}^T,\end{aligned}$$
所以
$$Z = (\mathbf{d}\mathbf{d}^T)^{-1}. \tag{5.4.18}$$
因而有
$$\begin{aligned}Q^TY + \mathbf{p}^TZ &= -Q^T(Q^+)^T\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1} + \mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1} \\ &= (I - Q^+Q)^T\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1} = (I - C^+C)^T\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1} \\ &= \mathbf{d}^T(\mathbf{d}\mathbf{d}^T)^{-1} = \mathbf{d}^+,\end{aligned} \tag{5.4.19}$$
$$\begin{aligned}&(Q^TW + \mathbf{p}^TY^T)D^+ = [Q^T(QQ^T)^{-1} - Q^T(QQ^T)^{-1}Q\mathbf{p}^TY^T \\ &- \mathbf{p}^TZ\mathbf{p}Q^+]D^+ = [Q^+ + Q^+Q\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}Q^+ - \mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}Q^+]D^+ \\ &= Q^+D^+ + Q^+Q\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}Q^+D^+ - \mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}Q^+D^+ \\ &= C^+ + C^+C\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}C^+ - \mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}C^+ \\ &= C^+ - (I - C^+C)\mathbf{p}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}C^+ \\ &= C^+ - \mathbf{d}^T(\mathbf{d}\mathbf{d}^T)^{-1}\mathbf{p}C^+ = C^+ - \mathbf{d}^+\mathbf{p}C^+.\end{aligned} \tag{5.4.20}$$
由(5.4.17)、(5.4.19)、(5.4.20)可得
$$\begin{aligned}A^+ &= \begin{bmatrix} Q \\ \mathbf{p} \end{bmatrix}^+ \begin{bmatrix} D^+ & O \\ O & 1 \end{bmatrix} = [Q^TW + \mathbf{p}^TY^T \; Q^TY + \mathbf{p}^TZ]\begin{bmatrix} D^+ & O \\ O & 1 \end{bmatrix} \\ &= [(Q^TW + \mathbf{p}^TY^T)D^+ \; Q^TY + \mathbf{p}^TZ] \\ &= (C^+ - \mathbf{d}^+\mathbf{p}C^+ \; \mathbf{d}^+).\end{aligned}$$

引理5.4.3 设 I 是 n 阶单位阵，$\mathbf{x} = (x_1, x_2, \cdots, x_n)$ 为 n 维行向量，则
$$(I + \mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T = \mathbf{x}^T/(1 + \|\mathbf{x}\|^2). \tag{5.4.21}$$

证明 因为据Householder变换 Q，可将任一列向量 $\mathbf{x}^T$ 变成
$$Q\mathbf{x}^T = \|\mathbf{x}\|\mathbf{e},$$

其中 e 为单位向量。因此，

$$Q(I+x^Tx)^{-1}x^T = Q(I+x^Tx)^{-1}Q^TQx^T$$

$$= [Q(I+x^Tx)^{-1}Q^T]Qx^T = [Q(I+x^Tx)Q^T]^{-1}Qx^T$$

$$= [I+(Qx^T)(Qx^T)^T]^{-1}Qx^T$$

$$= [I+\|x\|^2ee^T]^{-1}\|x\|e$$

$$= \frac{\|x\|}{1+\|x\|^2}e = \frac{Qx^T}{1+\|x\|^2}, \tag{5.4.22}$$

在(5.4.22)的两边乘 Q^T，即得

$$(I+x^Tx)^{-1}x^T = x^T/(1+\|x\|^2).$$

定理5.4.1 设$C\in R^{k\times n}$，$p\in R^{1\times n}$，且

$$A=\begin{bmatrix} C \\ p \end{bmatrix},$$

则

$$A^+ = [C^+ - fpC^+ \ f], \tag{5.4.23}$$

$$f=\begin{cases} d^+ & \text{当}d\neq o\text{时}; \\ C^+(C^T)^+p^Th & \text{当}d=o\text{时}; \end{cases}$$

$$d=p(I-C^+C),\ h=(1+pC^+(C^T)^+p^T)^{-1}.$$

证明 当$d\neq o$ 时，设$C=DQ$为 C 的秩分解。因为 $t+1\leqslant n$，否则$C^+C=I$与$d\neq o$矛盾；而且p与Q的行向量线性无关，否则，必有 $a\in R^{1\times t}$ 使$p=aQ$，于是

$pC^+C=p(Q^+D^+DQ)=p(Q^+Q)=aQQ^+Q=aQ=p$，这也和 $d\neq o$矛盾，因此$\begin{bmatrix} Q \\ p \end{bmatrix}$是行满秩阵，由此得到了$A$的一个秩分解

$$A=\begin{bmatrix} C \\ p \end{bmatrix}=\begin{bmatrix} D & O \\ O & 1 \end{bmatrix}\begin{bmatrix} Q \\ p \end{bmatrix}.$$

根据引理5.4.2得到

$$A^+=\begin{bmatrix} Q \\ p \end{bmatrix}^+\begin{bmatrix} D^+ & O \\ O & 1 \end{bmatrix}=(C^+-d^+pC^+d^+). \tag{5.4.24}$$

当$\boldsymbol{d}=\boldsymbol{o}$时，$\boldsymbol{p}-\boldsymbol{p}C^+C=\boldsymbol{o}$，即$\boldsymbol{p}=\boldsymbol{p}C^+C$，故

$$A=\begin{bmatrix} C \\ \boldsymbol{p} \end{bmatrix}=\begin{bmatrix} C \\ \boldsymbol{p}C^+C \end{bmatrix}=\begin{bmatrix} I \\ \boldsymbol{p}C^+ \end{bmatrix}C=BC.$$

设

$$B^+=\begin{bmatrix} I \\ \boldsymbol{p}C^+ \end{bmatrix}^+=(W,\boldsymbol{u}).$$

并作下三角阵

$$(B,\boldsymbol{v})=\begin{bmatrix} I & O \\ \boldsymbol{p}C^+ & 1 \end{bmatrix}, \tag{5.4.25}$$

其中$\boldsymbol{v}^T=(O^T,\ 1)$，显然$\boldsymbol{v}^T$与$B^T$的行线性无关，且

$$\begin{bmatrix} B^T \\ \boldsymbol{v}^T \end{bmatrix}=\begin{bmatrix} I & O \\ O & 1 \end{bmatrix}\begin{bmatrix} B^T \\ \boldsymbol{v}^T \end{bmatrix},$$

由引理5.4.2证明的结果可以得到

$$\begin{bmatrix} B^T \\ \boldsymbol{v}^T \end{bmatrix}^+=[(B^T)^+-(\boldsymbol{d}_1{}^T)^+\boldsymbol{v}^T(B^T)^+(\boldsymbol{d}_1{}^T)^+]$$

或

$$[B\boldsymbol{v}]^+=\begin{bmatrix} B^+-B^+\boldsymbol{v}\boldsymbol{d}_1^+ \\ \boldsymbol{d}_1^+ \end{bmatrix}, \tag{5.4.26}$$

其中

$$\boldsymbol{d}_1=(I-BB^+)\boldsymbol{v}=[I-\begin{bmatrix} I \\ \boldsymbol{p}C^+ \end{bmatrix}(W,\ \boldsymbol{u})]\begin{bmatrix} O \\ 1 \end{bmatrix}$$

$$=\begin{bmatrix} -\boldsymbol{u} \\ 1-\boldsymbol{p}C^+\boldsymbol{u} \end{bmatrix}. \tag{5.4.27}$$

又由于

$$(B,\boldsymbol{v})^+=(B,\boldsymbol{v})^{-1}=\begin{bmatrix} I & O \\ -\boldsymbol{p}C^+ & 1 \end{bmatrix}, \tag{5.4.28}$$

比较(5.4.26)与(5.4.28)可得

$$\boldsymbol{d}_1^+=(-\boldsymbol{p}C^+,\ 1). \tag{5.4.29}$$

又由(5.4.27)可得

$$\boldsymbol{d}_1^T\boldsymbol{d}_1 = \boldsymbol{v}^T(I - BB^+)\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{d}_1$$
$$= (0,1)\begin{bmatrix} -\boldsymbol{u} \\ 1 - pC^+\boldsymbol{u}\end{bmatrix} = 1 - pC^+\boldsymbol{u}.$$

故

$$\boldsymbol{d}_1^+ = (\boldsymbol{d}_1^T\boldsymbol{d}_1)^{-1}\boldsymbol{d}_1^T = (-(1 - pC^+\boldsymbol{u})^{-1}\boldsymbol{u}^T,\ 1). \qquad (5.4.30)$$

比较(5.4.30)与(5.4.29)得到

$$pC^+ = (1 - pC^+\boldsymbol{u})^{-1}\boldsymbol{u}^T. \qquad (5.4.31)$$

解(5.4.31)得

$$(pC^+)^T - (pC^+)^T(pC^+\boldsymbol{u}) = \boldsymbol{u},$$
$$(pC^+)^T = (I + (pC^+)^T pC^+)\boldsymbol{u},$$
$$\boldsymbol{u} = [I + (pC^+)^T pC^+]^{-1}(pC^+)^T, \qquad (5.4.32)$$

因为 pC^+ 是 n 维行向量，根据引理 5.4.3，(5.4.32)可以写成

$$\boldsymbol{u} = [1 + pC^+(pC^+)^T]^{-1}(pC^+)^T. \qquad (5.4.33)$$

又由于B 为列满秩阵，据定理5.3.3的推论 1 可得

$$I = B^+B = (W, \boldsymbol{u})\begin{bmatrix} I \\ pC^+\end{bmatrix} = W + \boldsymbol{u}pC^+,$$

即

$$W = I - \boldsymbol{u}pC^+. \qquad (5.4.34)$$

由于

$$A = BC$$

且

$$B = \begin{bmatrix} I \\ pC^+\end{bmatrix}.$$

据引理5.4.1可得

$$A^+ = C^+B^+ = C^+(W, \boldsymbol{u}) = C^+(I - \boldsymbol{u}pC^+, \boldsymbol{u})$$
$$= (C^+ - C^+\boldsymbol{u}pC^+, C^+\boldsymbol{u}).$$

令

$$\boldsymbol{f} = C^+\boldsymbol{u} = C^+[1 + pC^+(pC^+)^T]^{-1}(pC^+)^T$$
$$= C^+(pC^+)^T[1 + pC^+(pC^+)^T]^{-1} = C^+(pC^+)^T h, \qquad (5.4.35)$$

即

$$A^{+}=(C^{+}-fpC^{+},\ f).$$

根据定理5.4.1，矩阵

$$A=\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \qquad (\text{其中 } a_i \text{ 为 } n \text{ 维行向量})$$

的广义逆矩阵 A^{+}，可以使用如下递推算法。

算法5.4.1 本算法可以用依次计算 $A_1=a_1$，

$A_2=\begin{bmatrix} A_1 \\ a_2 \end{bmatrix}, \cdots, A_m=\begin{bmatrix} A_{m-1} \\ a_m \end{bmatrix}$ 的广义逆矩阵的方法，最后算得 A

的广义逆矩阵 A^{+}，其具体步骤如下。

1) 置 A_0^{+}，A_0 分别是 $R^{n\times 1}$ 和 $R^{1\times n}$ 中的零阵；

2) 对 $k=1,2,\cdots,m$，

  2.1) 置 $b_k=a_kA_{k-1}^{+}$

    2.1.1) 置 $d=a_k-b_kA_{k-1}$，

  2.2) 如果 $d\neq 0$，转2.4)，

  2.3) 置 $h=(1+b_kb_k^T)^{-1}$，

    2.3.1) 置 $f=A_{k-1}^{+}b_k^Th$，

  2.4) 置 $f=d^T(dd^T)^{-1}$，

  2.5) 置 $A_k^{+}=[A_{k-1}^{+}-fb_k,\ \ f]$，

  2.6) NEXT k。

3) 输出结果 A_m^{+}。

本算法略去了计算过程中的存贮技巧和计算细节。

例5.4.1 应用递推算法计算

$$C=\begin{pmatrix}-1&2&1\\-1&2&1\\0&3&2\end{pmatrix}$$

的广义逆矩阵C^+。

解 设$C_1=[-1,2,1]$

则

$$c_1^+=c_1^T(c_1c_1^T)^{-1}=[-1,2,1]^T\cdot\frac{1}{6}=\frac{1}{6}\begin{pmatrix}-1\\2\\1\end{pmatrix}.$$

设

$$C_2=\begin{bmatrix}c_1\\-1\ \ 2\ \ 1\end{bmatrix},\qquad \text{求}C_2^+.$$

由于

$$d=[-1,2,1]\left[I-\frac{1}{6}\begin{pmatrix}1&-2&-1\\-2&4&2\\-1&2&1\end{pmatrix}\right]$$

$$=[-1,2,1]\begin{pmatrix}\frac{5}{6}&\frac{2}{6}&\frac{1}{6}\\\frac{2}{6}&\frac{2}{6}&-\frac{2}{6}\\\frac{1}{6}&-\frac{2}{6}&\frac{5}{6}\end{pmatrix}=[0,0,0],$$

所以，取

$$f=c_1^+(c_1^+)^Tp^Th=c_1^+(c_1^+)^Tp^T[1+pc_1^+(c_1^+)^Tp^T]^{-1}$$

$$=c_1^+(pc_1^+)^T[1+pc_1^+(pc_1^+)^T]^{-1}=\frac{1}{6}\begin{pmatrix}-1\\2\\1\end{pmatrix}\cdot\frac{1}{2}=\frac{1}{12}\begin{pmatrix}-1\\2\\1\end{pmatrix},$$

$$C_2^+=\left(\frac{1}{6}\begin{pmatrix}-1\\2\\1\end{pmatrix}-\frac{1}{12}\begin{pmatrix}-1\\2\\1\end{pmatrix},\ \frac{1}{12}\begin{pmatrix}-1\\2\\1\end{pmatrix}\right)=\frac{1}{12}\begin{pmatrix}-1&-1\\2&2\\1&1\end{pmatrix},$$

设 $C_3=\begin{bmatrix} C_2 \\ 0\ \ 3\ \ 2\end{bmatrix}$，　求$C_3^+$。

由于

$$d=(0,3,2)\left(I-\frac{1}{12}\begin{pmatrix}-1 & -1\\ 2 & 2\\ 1 & 1\end{pmatrix}\begin{pmatrix}-1 & 2 & 1\\ \\ -1 & 2 & 1\end{pmatrix}\right)$$

$$=(0,3,2)\begin{pmatrix}\frac{5}{6} & \frac{1}{3} & \frac{1}{6}\\ \frac{1}{3} & \frac{1}{3} & -\frac{1}{3}\\ \frac{1}{6} & -\frac{1}{3} & \frac{5}{6}\end{pmatrix}=\left(\frac{4}{3},\frac{1}{3},\frac{2}{3}\right)\neq 0,$$

所以，取

$$f=d^+=\begin{pmatrix}\frac{4}{3}\\ \frac{1}{3}\\ \frac{2}{3}\end{pmatrix}\left(\left(\frac{4}{3},\frac{1}{3},\frac{2}{3}\right)\begin{pmatrix}\frac{4}{3}\\ \frac{1}{3}\\ \frac{2}{3}\end{pmatrix}\right)^{-1}=\frac{1}{7}\begin{pmatrix}4\\ 1\\ 2\end{pmatrix},$$

$$C^+=C_3^+=(C_2^+-fpC_2^+,\ f)=\left(\frac{1}{12}\begin{pmatrix}-1 & -1\\ 2 & 2\\ 1 & 1\end{pmatrix}-\frac{2}{21}\begin{pmatrix}4 & 4\\ 1 & 1\\ 2 & 2\end{pmatrix}\ \frac{1}{7}\begin{pmatrix}4\\ 1\\ 2\end{pmatrix}\right)$$

$$=\frac{1}{28}\begin{pmatrix}-13 & -13 & 16\\ 2 & 2 & 4\\ -3 & -3 & 8\end{pmatrix}.$$

线性最小二乘法问题，曾经一度只用法方程组$A^TAx=A^Tb$来求解。但是，由于方程组的条件数之间存在着$\mathrm{Cond}(A^TA)=(\mathrm{Cond}(A))^2$关系，所以当$\mathrm{Cond}(A)$较大时，法方程组就变得非常“病态”，用这种方法便难于求得满意的解了。此时，为了

避免处理非常病态的方程组，可以采用正交化方法和镜像变换法，“直接”对矛盾方程组进行求解。

还须注意，当矩阵A的秩小于n时，最小二乘解不唯一，从而需要求模为最小的最小二乘解。为此，需要用到奇异值分解，这种分解的难点是，确定(5.3.2)中哪些u_i应视为0。Golub与Reinsch于1970年提出了一种数值上十分稳定的分解方法，再加上奇异值的稳定性(见习题2)，意味着这种方法是较为可靠的。我们可以把这样得到的解用广义逆矩阵来表示。目前广义逆矩阵已有很多计算办法，它可以不通过奇异值分解而直接进行计算，由于篇幅有限，在此，我们仅介绍了一种递推算法。

第五章习题

5.1　证明　(1) $(AA^+)^2=AA^+$; (2) $(A^+A)^2=A^+A$; (3) $(A^+)^+=A$。

5.2　设$\sigma_i(A)$和$\sigma_i(A+E)$分别为矩阵A和$A+E$的奇异值，证明$\sigma_i(A)-\|E\|\leqslant\sigma_i(A+E)\leqslant\sigma_i(A)+\|E\|$。

5.3　给出应用奇异值分解求解齐次线性方程组的方法。

5.4　证明

若A是列满秩阵，则$(A^TA)^{-1}A^T=A^+$;

若A是行满秩阵，则$A^T(AA^T)^{-1}=A^+$。

5.5　证明定理5.3.5中的等式$X=XAX$和$XA=(XA)^T$成立。

5.6　证明等式(5.4.15)。

5.7　设$A\in R^{m\times n}$是列满秩阵，$C\in R^{l\times n}$是行满秩阵，导出一个在约束条件$Cx=d$下，使$\|b-Ax\|$达到最小的最小二乘解的算法，其中b和d均为已知的列向量。

5.8　求

$$\begin{cases}y_1+y_2+y_3=1,\\2y_1-y_2=0\end{cases}$$

的模为最小的解。

5.9　证明　在A的奇异值分解$A = UDV^T$中，V的行和U的列分别是A^TA和AA^T的特征向量。

5.10 利用对称分块矩阵求逆法，求

$$A = \begin{pmatrix} 3 & 0 & -1 & 0 \\ 0 & 3 & 0 & -1 \\ -1 & 0 & 3 & -1 \\ 0 & -1 & -1 & 3 \end{pmatrix}$$

的逆矩阵A^{-1}。

5.11 矛盾方程组$Ax = b$的最小二乘解(5.3.8)与通解(5.3.19)是否等价？为什么？

5.12 利用逆推算法计算

$$A = \begin{pmatrix} 1 & 3 & 1 & 1 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

的广义逆矩阵A^+。

5.13 给定数据表

x_i	0.000	1.445	2.890	4.335	5.780
y_i	1.8419	2.9633	18.2360	98.7410	529.2178

求形如 $y = ae^{bx}$ 函数的最小二乘方拟合。

5.14　求 $Ax = b$ 的模为最小的最小二乘解，其中

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 0 & 3 & -2 \\ 2 & 4 & -3 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}.$$

参　考　书

[1] Stewart. G. "*Introduction to Matrix Computation*" Academic Press. New york, 1973.

[2] 曹志浩等编，矩阵计算和求方程根。

[3] 党诵诗编，矩阵论及其在测绘中的应用，测绘出版社，1980。

第六章　稀疏矩阵

有限元方法、结构分析、网络分析及数学规划中的许多问题，都将归结为求解大型线性稀疏方程组的问题：

$$Ax = b. \tag{6.0.1}$$

这种方程组的特点是：矩阵A的阶数n很大，但其非零元素的个数τ所占的比例却很小。若以τ/n^2表示矩阵A的稀疏度，通常当τ/n^2小于0.05～0.25时，即认为A是稀疏的。

稀疏矩阵技术，就是针对矩阵的稀疏性来建立相应算法的一种技术。其基本出发点是，只让非零的信息参与运算和存贮，这样往往能在存贮量和机器时间的节省带来巨大的收益。限于篇幅，本章只着重讨论用直接法求解稀疏线性方程组的问题。

§1　稀疏矩阵的存贮

1.1　等带宽矩阵的存贮

对于矩阵A，如果存在正整数m，使

$$a_{ij} = 0,\quad |i-j|>m \qquad (i,j=1,\cdots,n,\ ) \tag{6.1.1}$$

则称A是**等带宽**的，称$2m+1$为其带宽，m为半带宽，把满足$|i-j|\leqslant m$的所有的(i,j)元素的集合叫做A的**带形区域**。

当对方程(6.0.1)进行不带行交换的高斯消去法时，在带形区域之外将不会产生非零元素。这时，我们只需把A的带形区域部分的元素存入机器。

注意到矩阵A的带形区域的形状，见图6.1的阴影部分，它只需补充少数几个元素就可当作二维数组来存放。

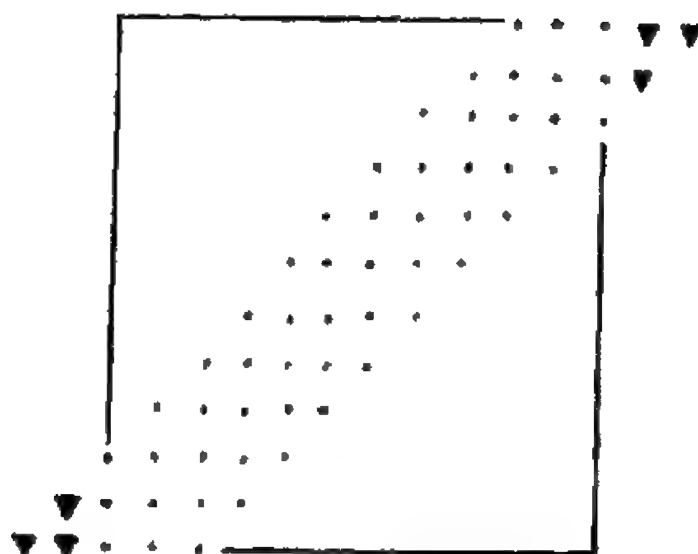


图6.1　等带宽矩阵

图中画有"▼"的元素，其数值可以任意给定。把得到的二维数组记为 $L[1:n,1:2m+1]$；用 i、j 表示元素 a_{ij} 在 A 中的行、列，用 i',j' 表示 a_{ij} 在 L 中的行列，则

$$
\begin{aligned}
&i'=i,\ \ j'=j-i+m+1,\\
&i=1,2,\cdots,n,\\
&\max(1,i-m)\leqslant j\leqslant\min(n,i+m). \qquad (6.1.2)
\end{aligned}
$$

(6.1.2)指出了 A 在带形区域中任意一个元素的位置与该元素在 L 中的位置的检索关系。所谓矩阵的压缩存贮，主要就是要建立这样的位置检索关系。当 A 是对称阵时，等带宽矩阵可以只存贮其下三角阵部分，设将 a_{ij} 存在二维数组 $L[1:n,1:m+1]$ 的 (i,j) 处，则

$$
\begin{aligned}
&i'=i,\ \ j'=i-j+1,\ \ j\leqslant i,\\
&i=1,2,\cdots,n,\ \ j=1,\cdots,\max(1,i-m).
\end{aligned}
$$

1.2　变带宽矩阵的存贮

如果 A 是对称阵，我们把 A 的第 i 行上对角线左侧离对角线最远的非零元素到同行对角元间的距离记为 β_i，称为第 i 行

上的半带宽。当A是正定对称阵时，由于可以进行 Cholesky 分解

$$A=LL^T,$$

其中L是下三角阵，可以证明，L阵在第i行上的半带宽一定与β_i相等。这样，当把A的三角分解的输出L的元素放在A的下三角阵中对应的元素上时，则A和L的存贮量要求将是完全一致的。

A的变带宽存贮是这样实现的：逐行把A的下三角部分的每一行的元素（包括半带宽范围内的每一个零元和非零元），依次存入一个一维数组$L[1:s]$中，其中

$$s=\sum_{i=1}^{n}(\beta_i+1), \tag{6.1.3}$$

$$0\leqslant\beta_i\leqslant i-1,\quad i=1,\cdots,n.$$

令A的第i个对角元在L数组中的地址为$d[i]$，则$a_{ij}\ (j\leqslant i)$在L中的地址l可表为

$$l=d[i]+j-i, \tag{6.1.4}$$

$$d[i]=\sum_{k=1}^{i}(\beta_k+1). \tag{6.1.5}$$

反之，由L数组中的第l个数，也可以求出该数在原矩阵A中的行号i和列号j：

$$i=\max_k(l>d[k])+1,$$

$$j=l-d[i-1]+i-\beta_i=l-d[i]+i. \tag{6.1.6}$$

可见，在组织对称阵A的变带宽输入时，应输入两个一维数组$L[1:s]$和$D[1:n]$，前者是逐行存放A在各条半带宽上的元素的，后者是存放A的对角元在L中的位置的。

如图 6.2，设各有关数据为：$n=9$，$\beta_1=0$，$\beta_2=1$，$\beta_3=2$，$\beta_4=1$，$\beta_5=0$，$\beta_6=3$，$\beta_7=2$，$\beta_8=2$，$\beta_9=3$；$s=23$，$d_1=1$，$d_2=3$，

$d_3=6, d_4=8, d_5=9, d_6=13, d_7=16, d_8=19, d_9=23$; a_{76} 在 L 中的地址为：$l=d_7+6-7=15$，L 中的第 12 个数的行号 i、列号 j 各为：

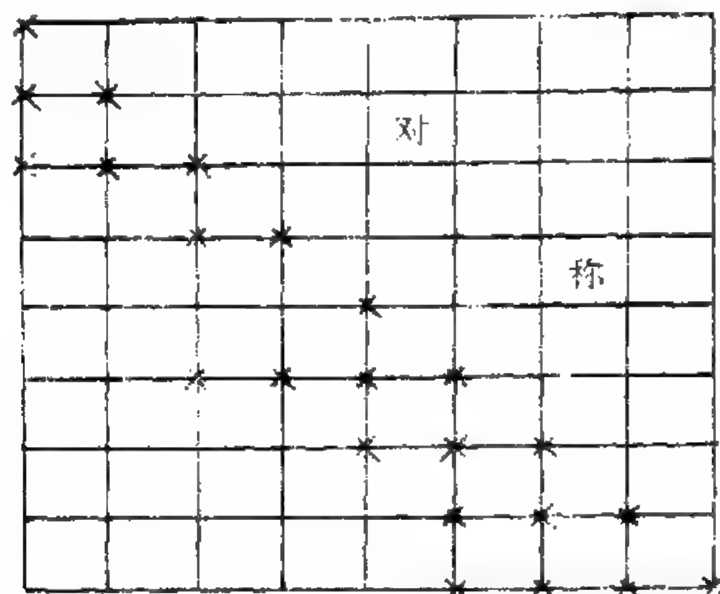


图6.2 对称变带宽矩阵

$$i=\max_k(12>d[k])+1=5+1=6.$$

$$j=12-d_6+6=5.$$

1.3 压缩存贮法

当矩阵 A 中的非零元素的分布十分零乱时，可用下述的压缩存贮法进行存贮。

在内存中指定 n 个连续编号的单元，称为列首址表，记为 $BC[1:n]$，$BC[j]$ 的内容被规定为 A 的第 j 列第 1 个非零元素的信息条首址表所谓一个非零元素 a_{ij} 的信息条，是指由三个连续编号的单元组成的成组单元，三个连续单元中的内容按图 6.3规定。

信息条首址	第二地址	第三地址
元素行号i	元素 a_{ij}	下一非零元信息条首址或零

图6.3 信息条内容

如图，如果在第 j 列的 a_{ij} 元素之下还有后续的非零元，就在第三地址的内容中填写下一个非零元的信息条首址；如果 a_{ij} 是第 j 列中最后一个非零元，就在第三地址中写零。这样，当知道了一个元素的信息条首址之后，不仅同时能确定该元素的值，而且连下一个非零元的信息条首址也知道了。由于同列的非零元的所有的信息条是一一连接的，所以这种存贮方法又有连接表法之称。

当知道元素的行号 i 和列号 j 时，查找元素 a_{ij} 的步骤应当是：

1）由 j 找 $BC[j]$，设其内容为：

$$(BC[j]) = d,$$

d 是第 j 列第 1 个非零元的信息首址。

2）如果 $(d) = i$ 则 $(d+1) = a_{ij}$，于是检索 a_{ij} 的工作结束；

3）如果 $(d) \neq i$，则 $(d+2) \Rightarrow d$，并转2)；只要元素的行、列信息没有填错，总有一次会使 2) 成立。

设矩阵 A 的非零元的总数为 τ，则存贮全部信息所需的存贮单元总数为 $3\tau+n$，当 $3\tau+n \ll n^2$ 时，这对存贮的节约是很有利的。

当用直接法求解方程组 (6.0.1) 时，原矩阵中原来是零的元素，在消元过程中可能变成非零的元素，原来是非零的元素又可能消元为零，于是在信息条的"链条"中，必须讨论"增添"和"删除"一个元素的问题。上述存贮方法能方便地实现这种要求。

首先，假定在内存中有一个有足够容量的空白单元区，其首址为 L，它是由三个连续编号的单元组成的空白条的"链条"组成的，空白条与信息条的区别在于：空白条的第一、第二地址的内容可以是任意的，但第三地址的内容必须填写下一个空

白条的首址．而空白区的首址 L 存放在一个固定的地址 S 中，即令

$$(S) = L.$$

于是当要在矩阵的第 j 列中增加一个元素时，例如，设在第 j 列中，相邻的两个非零元 $a_{i_1 j}$、$a_{i_2 j}$ 已经分别放在首址为 d_1、d_2 的相连接的信息条之中，其中 $i_2 - i_1 \geqslant 2$，

d_1	d_1+1	d_1+2	…	d_2	d_2+1	d_2+2	…
i_1	$a_{i_1 j}$	d_2		i_2	$a_{i_2 j}$	p	

今欲在第 j 列中的 $a_{i_1 j}$、$a_{i_2 j}$ 之间插入一元素 $a_{ij}(i_1<i<i_2)$，应按下列步骤进行：

1）求 $(s) = L$，找到空白区首址；

2）$(L+2) \Rightarrow s$，把下一空白条首址作为空白区首址；

3）在 $L, L+1, L+2$ 中，分别记入：i, a_{ij}, d_2，即“征用”原空白区的第 1 个空白条为信息条以记录 a_{ij} 及其行号，并把新信息条的下端接好；

4）把 L 记入 d_1+2，即把新信息条的上端也接好．

当 a_{ij} 是第 j 列中的第 1 个要增新的非零元或最末一个要增新的非零元时，亦可仿此进行连接．

当要从第 j 列中“删除”一个元素时，例如，设 $a_{i_1 j}$，$a_{i_2 j}$，$a_{i_3 j}$ 已经分别存放在以 d_1、d_2、d_3 为首址的互相连接的信息条之中，即

d_1	d_1+1	d_2+2	…	d_2	d_2+1	d_2+2	…	d_3	d_3+1	d_3+2
i_1	$a_{i_1 j}$	d_2		i_2	$a_{i_2 j}$	d_3		i_3	$a_{i_3 j}$	p

今欲删除以 d_2 为首址的信息条，为使被删除的信息条能“退役”到空白区中，以备将来征用，可按下列步骤实现：

1) $(d_2+2)\Rightarrow d_1+2$，这就完成了对 d_2 条的删除，

2) 由 $(s)=L$，求出 L，并令 $d_2\Rightarrow s$，$L\Rightarrow d_2+2$，即把删除掉的信息条作为空白区的第 1 个空白条，并使之与原有空白条连接起来。

还可设计出一些压缩存贮的方法。但使用上述连接表方法，却独具一种优点，即增删一个元素都很方便，特别，欲变换矩阵的两列时，例如，欲交换第 j_1 和 j_2 两列，只需把 $B[j_1]$ 和 $B[j_2]$ 的内容互相交换即可。这种思想还可很容易地推广到复稀疏结构的情形。

§2 消元次序

2.1 引言

在稀疏矩阵计算中，在设计或选用一种算法时，应尽可能使计算过程所需的最大存贮量和运算次数达于极小，为了达到这个目标，先介绍几个术语。

填充量 如图 6.4 所示的 7 阶矩阵中，用 × 表示非零元素，空白处为零元素。在进行消元过程的第一步时，应把第 1 列上三角化，此时第 1 列上的 (2,1)，(3,1)，(7,1) 应消元成零。与此同时，原来有些是零的元素如(2,3)，(2,6)，(3,6)，(7,3)，(7,6)却将由零变成非零。图中用⊗标出的地方，就是这些新增的非零的元素。在本例中，其个数是 5 个，称为第一步消元过程中产生的填充量。如以此

$$\begin{bmatrix} \times & & \times & & & \times & \\ \times & \times & \otimes & & & \otimes & \times \\ \times & & \times & & & \otimes & \\ & & & \times & & & \\ & & \times & & \times & & \\ & & & & & \times & \\ \times & & \otimes & & & \otimes & \times \end{bmatrix}$$

图6.4 填充量

数减去原第1列中的被消元成零的非零元个数，本例中应为 $5-3=2$，可称之为第1次消元时的填充增量。类似地，在第2次消元时也会出现新的填充量增量，而在整个消元过程中，填充量增量将有一个最大的积累。

值得指出的是：填充量的积累与主元选取的次序或矩阵的行列的排列次序有关。如图6.5，当按自然次序消元时，则在第一步消元时，填充量即达于最大；但若把第1行、列与最末的行列交换时，则整个消元过程中，填充量保持为零。下一节我们将介绍一些重排行、列次序的方法，以使填充量积累趋于极小。

$$\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & & & \\ \times & & \times & & \\ \times & & & \times & \\ \times & & & & \times \end{bmatrix} \qquad \begin{bmatrix} \times & & & & \times \\ & \times & & & \times \\ & & \times & & \times \\ & & & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix}$$

(a)　　　　　　　　(b)

图6.5　填充量与行列交换

主元容限　在利用列主元或全主元消去法把矩阵上三角化的过程中，例如，在做第1步列主元消去时，通常应把第1列中绝对值最大的元素换行到(1,1)的位置，然后进行消元。这样做的目的，一般是为了提高计算精度的。但是，如果当前的主元的选取恰好导致填充量增加最大，这就不利于内存的节省，于是宁愿选择一个绝对值不是最大但又不会引起填充量有过大增长的元素作主元。当然，作为主元的元素，其绝对值也不能过小，否则就会引起精度的大的丧失。人为地规定一个界限 $\varepsilon>0$，当矩阵的元素依模大于此数时，该元素就具有了当主元的资格，如果它引起的填充量也果然不大，就可把主元定下

来。所规定的这个界限 ε 即称为主元容限。这个限量可以经验地给以规定，但它应体现计算精度及存贮节省的统一要求。

主元容限的可供推荐的选取方法是：先求

$$a = \underset{i,j}{\mathrm{Max}}|a_{ij}|$$

并取 $\varepsilon = 10^{-4}a$。

主元容限的上述选取方法是以矩阵具有一定的均衡性为前提的，即矩阵的元素依模均不大于1，且每个行和每个列中都至少有一个元素依模不小于1/2。当矩阵非奇异时，通常总能找到两个对角阵 D_1 和 D_2，使 D_1AD_2 能成为均衡或接近均衡。

2.2 填充量极小化方法

在进行某一步消元时，该步中的填充量极小化问题是容易处理的。然而，即使每一步消元都保证了该步中的填充量极小化，也未必能保证得到全局的填充量极小化方案。要得到全局的使填充量积累最小的消元次序，原则上应比较所有可能的消元次序方案中实际填充量积累的大小，才能选出最好的方案。但是，所有可能的消元次序方案总数可达 $(n!)^2$ 之多！因此，现实的作法仍只需保证每步消元过程中的填充量极小就行了。但此种作法只能叫作局部的填充量极小化方法。

由于每一步消元过程中填充量极小化的讨论方法是一样的，所以我们只讨论第1步消元时的填充量极小化方法。

令 B 是 A 的布尔矩阵，其元素规定为

$$b_{ij} = \begin{cases} 1, & \text{当 } a_{ij} \neq 0 \text{ 时；} \\ 0, & \text{当 } a_{ij} = 0 \text{ 时。} \end{cases} \tag{6.2.1}$$

设在第1步消元过程中，以 a_{pq} 为主元。则 $a_{pq} \neq 0$，称 p 为主行，q 为主列。于是，在 (i,j) 元素处产生填充量的充要条件是

$$(a_{pj} \neq 0) \wedge (a_{iq} \neq 0) \wedge (a_{ij} = 0) \tag{6.2.2}$$

或

$$(b_{pj}=1)\wedge(b_{iq}=1)\wedge(b_{ij}=0) \tag{6.2.3}$$

或

$$e_p^T B e_j e_i^T \bar{B} e_j e_i^T B e_q = 1. \tag{6.2.4}$$

这里$\bar{B}$是B的共轭矩阵，它是把B中的0换成1，1换0后所成的矩阵。因此，当以a_{pq}为主元时，除p行和q列上的元素外，在其余各处产生填充量的总和为

$$g_{p,q}=\sum_{\substack{i,j\\ i\neq p,\ j\neq q}} e_p^T B e_j e_i^T \bar{B} e_j e_i^T B e_q. \tag{6.2.5}$$

注意，在上式中，当$i=p$或$j=q$时，由于

$$e_p^T B e_j e_p^T \bar{B} e_j = 0 \text{ 及 } e_i^T \bar{B} e_q e_i^T B e_q = 0,$$

故(6.2.5)中求和号下的$i\neq p$，$j\neq q$的限制可以去掉，得

$$\begin{aligned}
g_{p,q} &= \sum_{i,j=1}^{n} e_p^T B e_j e_i^T \bar{B} e_j e_i^T B e_q \\
&= \sum_{i,j=1}^{n} e_p^T B e_j e_j^T \bar{B}^T e_i e_i^T B e_q \\
&= e_p^T B \left(\sum_{j=1}^{n} e_j e_j^T\right) \bar{B}^T \left(\sum_{i=1}^{n} e_i e_i^T\right) B e_q \\
&= e_p^T B \bar{B}^T B e_q .
\end{aligned} \tag{6.2.6}$$

令

$$g_{\alpha\beta}=\min_{\substack{p,q\\ |a_{pq}|>\varepsilon}}\{e_p^T B \bar{B}^T B e_q\}, \tag{6.2.7}$$

则(α,β)应是第1次消元时产生的使填充量极小化的主元位置。这个方法是1967年由Tewarson提出的。其后，Markowitz又把此法加以简化。称

$$b_{pj} b_{iq}$$

为第1步消元时，如以(p,q)为主元，而在(i,j)处可能产生的填充量，由于

$$b_{pj} b_{iq} \geqslant b_{pj}(1-b_{ij}) b_{iq}$$

类似于(6.2.6)的推导，可得

$$\bar{g}_{p,q}=\sum_{\substack{i\neq p\\ j\neq q}} b_{pj}b_{iq}=e_p^T(B-I)VV^T(B-I)e_q$$

$$=e_p^T(B-I)M(B-I)e_q, \qquad (6.2.8)$$

这里用V表示分量全为1的向量，M表示元素全为1的矩阵。令

$$\bar{g}_{\alpha,\beta}=\min_{\substack{p,q\\ |a_{p,q}|\geq\varepsilon}}\{e_p^T(B-I)M(B-I)e_q\}, \qquad (6.2.9)$$

这就得到了 Markowitz 的填充量极小化的策略。

注意，由于(6.2.8)中$e_p^T(B-I)V$和$V^T(B-I)e_q$分别表示与元素a_{pq}同行或同列的不包含a_{pq}的非零元的总数，因此$\bar{g}_{p,q}$的计算十分容易。在实践中，此法已被广泛的采用。

当矩阵A为对称阵时，为了保证消元过程中流动矩阵$A^{(k)}$的对称性，通常应把主元选在对角线上，即于(6.2.9)中，应取$\alpha=\beta$，得

$$\bar{g}_{\alpha}=\min_{\substack{p\\ |a_{pp}|\geq\varepsilon}}\{V^T(B-I)e_p\}^2. \qquad (6.2.10)$$

2.3 计算量的极小化方法

在消元过程中，欲使总的计算量趋于极小，选取主元时也应有一定的策略．我们仍以第1次消元时的计算量分析为主。

当a_{pq}被选为主元时，其进行过程为：先作一次除法，用以求主元的倒数$1/a_{p,q}$，V^TBe_q-1次乘法以实现主元除主列，这是矩阵三角分解中需要存贮的信息；再作e_p^TBV-1次乘法，实现主元除主行，这是消元时所必须的信息；最后作

$$\sum_{\substack{i\neq p\\ j\neq q}}(e_p^TBe_j)(e_i^TBe_q)=(e_p^TBV-1)(V^TBe_q-1)$$

次乘法，这是消元时需要作的。于是，1次消元时的乘除法总量为：

$$\bar{g}_{pq} = 1 + (V^T B e_p - 1) + (e_p^T B V - 1) + (e_p^T B V - 1) \times (V^T B e_q - 1) = e_p^T B M B e_q. \qquad (6.2.11)$$

比较(6.2.8)与(6.2.11)，并引入

$$h_{pq}(\mu) = e_p^T (B - \mu I) M (B - \mu I) e_q, \quad (0 \leqslant \mu \leqslant 1) \qquad (6.2.12)$$

则有

$$h_{pq}(0) = \bar{g}_{pq}, \quad h_{pq}(1) = \bar{g}_{pq}. \qquad (6.2.13)$$

上式分别表示，当 a_{pq} 为主元时，欲使计算量极小应取 $\mu = 0$，而欲使可能的填充量极小应取 $\mu = 1$。于是 $h_{pq}(\mu)$ 应视为一次消元过程中，当 a_{pq} 为主元时，使计算量和填充量达于极小的带权平均值。而

$$\bar{h}_{\alpha,\beta}(\mu) = \min_{\substack{p,q \\ |a_{pq}| \geqslant \varepsilon}} h_{pq}(\mu) \qquad (6.2.14)$$

将是综合计算量的极小化和填充量的极小化的选取主元的策略。权因子的选取，取决于机器的硬件和软件，当计算速度成为计算过程中的主要矛盾时，宜取 μ 接近于零，反之，则应取 μ 接近于1。

2.4 带宽极小化方法

由于对称带状阵在带形区域内的零元在存贮时也要分配单元，当带宽较大而带形区域内的零元又较多时，常可利用交换行的相似变换使带宽减小。如图6.6(a)，对给定的对称6阶矩阵，由于半带宽$m = 6$，依等带宽的压缩形式存贮时需占用21个单元。但若将原矩阵的第1,2,3,4,5,6各行和各列依次排在第1,6,2,5,3,4各行和各列时则得如图6.6(b)中的稀疏结构，此时半带宽为 $m = 1$，仅需占用11个单元。

这就自然地提出下述问题：求排列阵 P，使

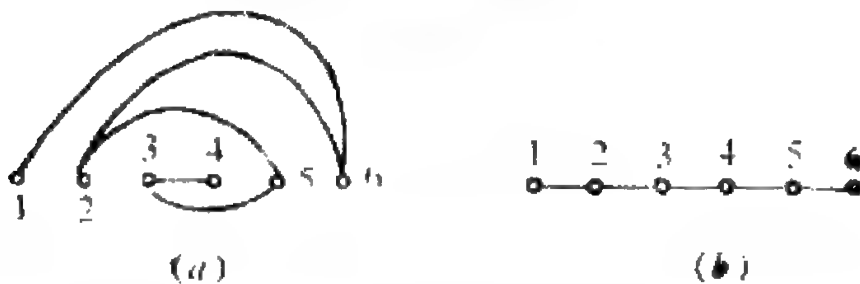
$$PAP^T$$

的带宽最小，其中A是对称正定矩阵。



(a)　　　　(b)

图6.6　带宽的约化



(a)　　　　(b)

图6.7　图6.6中的矩阵的无向图

我们将借用图论中的一些术语来讨论上述问题。

先设A为任意的n阶实矩阵，并于平面上任取互异的n个顶点$v_1, v_2, \cdots, v_n$。当矩阵A的元素$a_{ij} \neq 0$时，用一条边把顶点v_i和v_j连接起来，记为$\langle v_i, v_j \rangle$，称为边。当视$\langle v_i, v_j \rangle$与$\langle v_j, v_i \rangle$为同一条边时，称这种边为**无向边**，否则称为**有向边**，$\langle v_i, v_i \rangle$称为**自身路**，若记顶点的集合为$V = \{v_1, v_2, \cdots, v_n\}$，边的集合记为$E = \{\langle v_i, v_j \rangle\}$，则顶点与边的集合

$$G = \{V, E\} \tag{6.2.15}$$

称为图。

当所有的边均是无向边时，相应的图叫做**无向图**，反之叫做**有向图**。当A是对称阵时，通常我们只研究它的无向图。例如，前面已给的图6.6(a)，(b)中的矩阵的无向图为（见图6.7）：

如果在图G和图G'的顶点和边之间可以建立一一对应的关系，并且对应边上的对应点也是对应的，则称图G与图G'

同形，同形的图可视为同一个图。当A是对称阵时，显然其同形的无向图是唯一确定的，反之，同形的无向图也唯一地确定了对称阵的非零结构。当A是正定对称阵时，由于矩阵A的每个对角元均是正数，因此，其无向图上的每个点都有一条自身路，为了简化图面，今后在讨论正定对称阵的无向图时，我们将略去其所有的自身路。

交换对称阵的第i、j两行和两列，比较矩阵在交换前后的无向图，可知，只要在原有的无向图中，交换顶点v_i和v_j的编号，即可得到交换后的矩阵的无向图；由此推知，矩阵

$$PAP^T \tag{6.2.16}$$

乃是，在A的无向图上重新对顶点进行编号后的无向图的矩阵。于是求排列阵的问题就转变成如何对图的顶点进行编号，使PAP^T的带宽极小化的问题。

在图G中任意取定一个顶点v_i，称通过此点的边数为该点的**度**，这些边中不包括v_i本身的顶点叫做v_i的**邻点**，显然v_i的邻点的个数与v_i的度是相等的。

注意到矩阵A的第i行上的半带宽β_i，乃是v_i的编号减去v_i的邻点中比v_i编号小的最小的编号，当v_i的邻点中没有比v_i更小的编号时，则$\beta_i=0$。可见，寻求顶点的编号的方法，以使图的相应的矩阵有最小的带宽，其原则之一应当是，每个顶点v的编号与v的邻点的编号不宜相差过大。

Cuthill和Makee基于图的理论，用对顶点重新编号的方法，使矩阵的带宽减小，文献上称为CM方法，其后又为Rosen加以改进，发展成为RCM方法。目前，它已应用于许多大型结构的程序分析系统。

于图$G=\{V,E\}$的顶点的集合V中，任取度数最小的一个顶点v称为**根**，记

$$L_1 = \{v\} \tag{6.2.17}$$

称为**第一分层**，把 V 中与 v 相邻的点的集合（不包括 v）记为 L_2，称为第 2 个分层，一般地把与 L_{i-1} 相邻但不属于 L_{i-1} 与 L_{i-2} 的点的集合记为 L_i，叫做第 i 个分层；设 V 中的点一共可区分为 k 个分层，称

$$L_1, L_2, \cdots, L_k \tag{6.2.18}$$

为图的以 v 为根的一个分层结构。显然

$$L_i \cap L_j \neq 0 (i \neq j), \bigcup_{i=1}^{k} L_i = V \tag{6.2.19}$$

称每个分层 L_i 中的点数 w_i 为 l_i 的宽度，而把

$$w = \max_{1 \leq i \leq k} w_i \tag{6.2.20}$$

叫做**分层结构的宽度**，k 为**分层结构的深度**。根的选取，可能不是唯一的，选定以后，即将它编号成 1，然后依次给第二层中的点编号，在给第 i 层中的顶点编号时，应先编第 $i-1$ 层中度数最低的那些点的邻点。如这样的点不止一个，则应使第 i 层中编号大的点与编号小的点分别与第 $i-1$ 层编号大的点和编号小的点相邻。当各层均已编完后，一个排列阵 P 即已形成。以图6.7(a)为例，如以 v_1 为根，则按 CM 方法，各分层分别为：

$L_1 = \{v_1\}, L_2 = \{v_6\}, L_3 = \{v_2\}, L_4 = \{v_5\}, L_5 = \{v_3\}, L_6 = \{v_4\}$,

因为原来的点的编号 1, 2, 3, 4, 5, 6 应重排成 1, 3, 5, 6, 4, 2，即新的点的编号应为 1, 6, 2, 5, 3, 4，于是

$$P = (e_1, e_6, e_2, e_5, e_3, e_4).$$

当以 P, P^T 左乘、右乘于图 6.6(a)中的矩阵时，即得图 6.6(b)中所示的矩阵。

给图的顶点重新排序，使相应矩阵的带宽趋于极小，这是排序过程中所要追求的一个目标，此外，使矩阵的外形趋于极

小，则是排序中所要追求的另一个目标。

矩阵的外形是指

$$\beta=\sum_{i=1}^{n}(\beta_i+1), \tag{6.2.21}$$

其中 β_i 是矩阵的第 i 行上的半带宽。显然，当用变带宽方式存贮矩阵时，矩阵的外形的减小将有利于我们的讨论。

在 CM 方法的基础上如果使用反编号的方法，即把由 CM 方法得到的某顶点的编号 i 重新编号为 $n-i+1$，可达到使外形减小的目的。例如，以图 6.8 为图的矩阵，其外形为 109；而用反排序方法即 RCM 方法排序时，外形化为97。

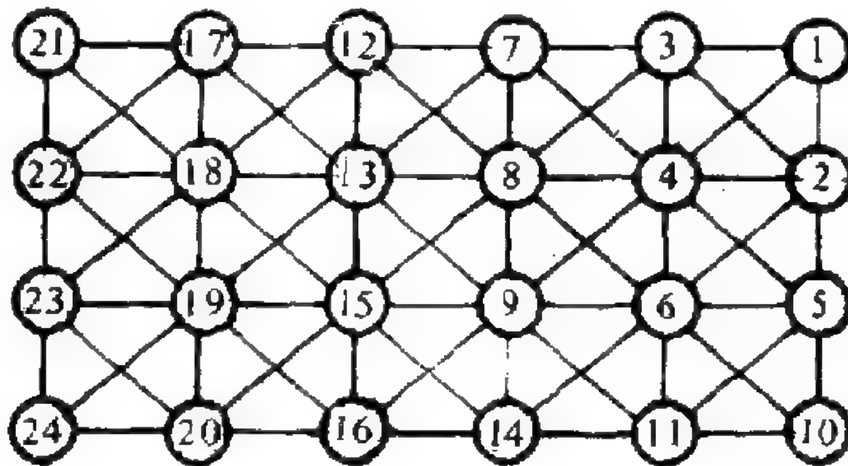


图6.8　用 CM 方法排序

RCM 方法也还存在着可以改进的地方。首先，RCM 方法需对所有度数较低的点进行搜索，逐个计算分层结构，每次计算中都包括求一个排列矩阵 P 和 PAP^T 的带宽，这些工作量是比较大的。为了减少这种搜索的次数，Gibbs, Poole, Stochmeyes 等人基于一种几何直觉，提出一种伪直径概念，设计了另一种求分层结构的方法，称为 GPS 方法。在某些情况下，使用这种方法确实显露了它的优点。

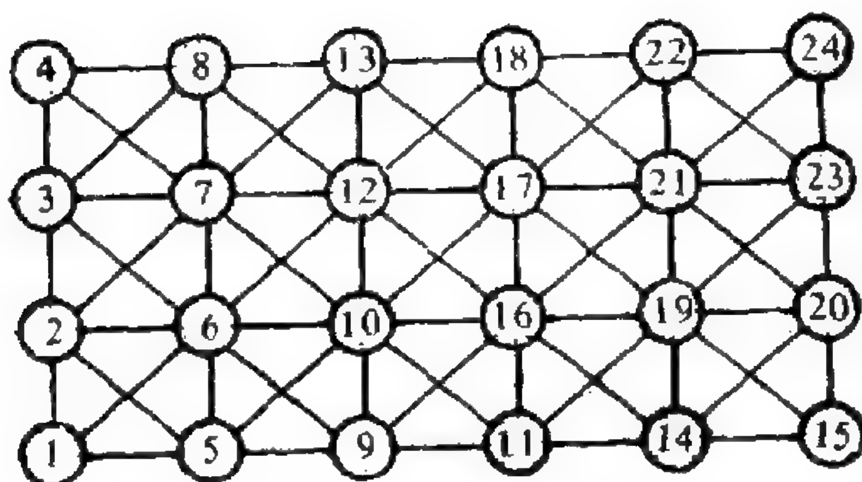


图 6.9 用 RCM 方法排序

几何直觉的启示是，分层结构的深度越大，则各分层结构的平均宽度将越小，注意 L_i 分层上一个点 v 的编号 r 减去 L_{i-1} 分层上与 v 点相邻且编号小于 r 的最小编号的号码就是矩阵 A 的第 i 行上的半带宽，这个值的上界必定不超过 L_i, L_{i-1} 的宽度之和，所以分层结构的深度大，矩阵的带宽将平均地越小。设以 v 为根的分层结构是图的具有最大深度 k 的一个分层结构，我们把从 v 到第 k 个分层上任意一点 u 之间的最短路径叫做图的一个直径，其长为 k，而 u、v 两点即称为直径的端点。显然，要求一个图的一条直径及其端点，理论上应进行大量的搜索，这是不现实的，简单而实用的作法是：

1）取任意一个度数最低的点 v；

2）以 v 为根做分层结构：$L_1, L_2, \cdots, L_k$，

3）在 L_k 上依度数增加的次序选点 s，以 s 为根做分层结构，如果深度不增加且 s 已经选完则转4)，如未选完，继续做3)，如深度有增加，则 s 送到 v 并转2)；

4）L_k 上任意一点与 v 之间的最短路径都是 k，可当作是图的直径的近似值，特称之为伪直径。于 L_k 上找一点 u，使以 u 为根做的分层结构所对应的矩阵的带宽 $w(u)$ 满足

$$w(u) = \min_{s \in L_k} w(s). \tag{6.2.22}$$

例如，在图(6.8)中，取 $v=1$，则分层结构的各分层为：

$L_1 = \{1\}, L_2 = \{2, 3, 4\}, L_3 = \{5, 6, 7, 8, 9\},$

$L_4 = \{10, 11, 12, 13, 14, 15, 16\}, L_5 = \{17, 18, 19, 20\},$

$L_6 = \{21, 22, 23, 24\};$

这时 $k=6$，由于 $w(21) = w(24) = 7$，$w(22) = w(23) = 6$，取 $u = 23$，以 $u = 23$ 为根的分层结构的层为

$M_1 = \{23\}$，$M_2 = \{24, 20, 19, 18, 22\}$，

$M_3 = \{16, 15, 13, 12, 17, 21\}$，$M_4 = \{14, 9, 8, 7\}$，

$M_5 = \{11, 6, 4, 3\}$，$M_6 = \{10, 5, 2, 1\}$

反编 M_i 的层号，即记

$$M'_i = M_{k+1-i},\quad i = 1, 2, \cdots, k,$$

由于图中每个顶点都将分别属于分层结构 L 和分层结构 M' 的一个分层，因此，每个点各对应 L 分层的一个足码 i 和 M' 的一个足码 j，亦即对应一个足码对 (i, j)，如图 (6.10)，这个足码对的表称为图 6.8 的以 1, 23 为伪直径端点的**相关层表示**。如果把相关层表示图中，具有相关足码为 (i, i) 的那些点组成另一分层结构 N 的第 i 层，则此时

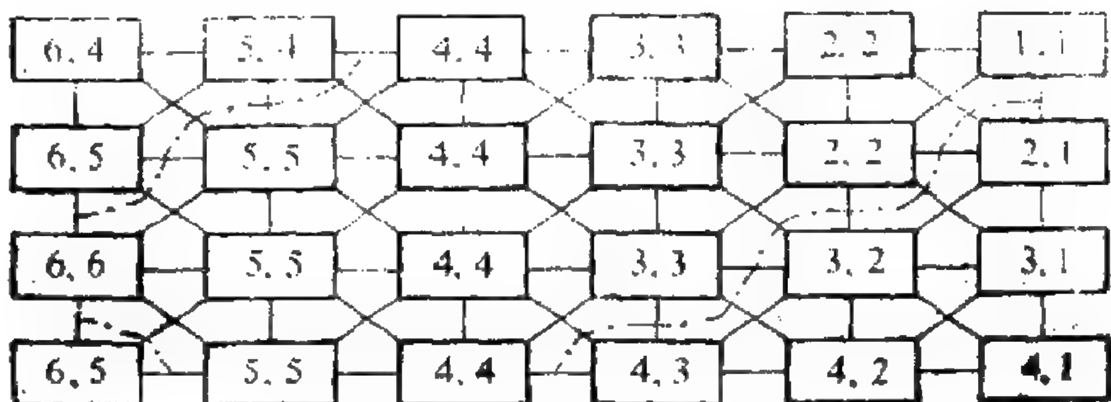


图6.10 图6.8的相关层表示

$N_1=\{1\}, N_2=\{3,4\}, N_3=\{7,8,9\},$

$N_4=\{12,13,15,16\}$, $N_5=\{20,19,18\}$, $N_6=\{23\}$.

它们表明，当以某一伪直径的端点为根时，N_i 中的点都应在同一层次。剩下还需把相关足码为 (i,j) $i\neq j$ 的点分别编入 N 的各层中，这些点在相关层表示图中，一般将分隔成互相隔离的几个连通的子集：

$$c_1, c_2, \cdots, c_t.$$

在上例中，$t=3$ 而：

$$c_1=\{2,6,14,5,11,10\},$$

$$c_2=\{21,17,22\},\ c_3=\{24\},$$

它们也是相关足码的集合：

$c_1=\{(2,1),(3,2),(4,3),(3,1),(4,2),(4,1)\}$,

$c_2=\{(6,4),(5,4),(6,5)\}$, $c_3=\{(6,5)\}$.

我们先来讨论 c_1 中的点的归属问题。在 c_1 中任选一个关连号码设为 $(4,3)$ 的点，如把它放入 N_4 层，就说这个点是按其第 1 足码决定其归属的；如放入 N_3 层，则是按第 2 足码决定归属的。因为 $(4,3)$ 这个点必定是 N_3 或 N_4 的邻点，所以决不应把它放入 N_3、N_4 以外的层中去，如果 c_1 中的点全部按第1足码决定其归属，则原来的 N_1, N_2, N_3, N_4 将扩充成：

$N_1=\{1\}$, $N_2=\{3,4,2\}$,

$N_3=\{7,8,9,6,5\}$, $N_4=\{12,13,15,16,14,11,10\}$.

按照前面提到的应使诸层最大宽度最小的原则，我们应按第 2 种方式来决定 c_1 中各点的归属，类似的讨论可施之于 c_2 和 c_3。当 c_1、c_2、c_3 中各点都已分配完毕后，即可逐层进行点的编号 $N_1, N_2, \cdots, N_6$ 各层中的点的划分情况及重新编号情况如图6.10和6.11相应的图，这时的带宽已缩减到 $\beta=5$，外形则已减到98。

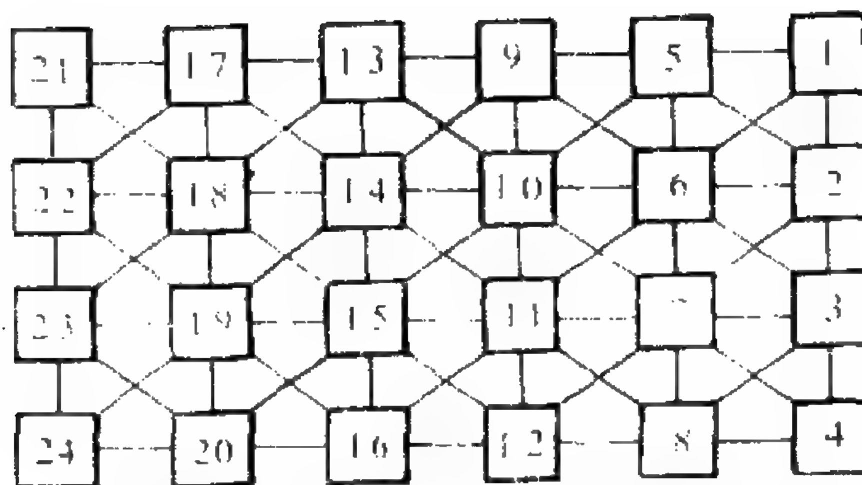


图6.11　按GPS方法决定顶点编号

§3　带状方程组解法

方程组的系数阵是带状阵的情形，在实用中较为普遍，且当联系到内存及计算量的节省时，有些实用性的技巧值得提出来，本节着重讨论这方面的问题。

3.1　等带宽矩阵的消去法

本段谈到的消去法是指列主元消去法，系数阵本身可以不必是对称的。

设

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \tag{6.3.1}$$

为所求方程组，系数阵 A 的带宽为 $2m+1$。当 A 以压缩形式存于二维数组 $L[1:n,1:2m+1]$ 时，记

$$L[i',j']=A[i,j],$$

$$i'=i,\ j'=j-i+m+1, \tag{6.3.2}$$

$$i=1,2,\cdots,n,\ \max(1,i-m)\leqslant j\leqslant\min(i+m,n)$$

当 $n=6, m=2$ 时，带形区域内的元素在矩阵 A，L 中的存贮情况见图6.12。

$$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} & & & \\ a_{21} & a_{22} & a_{23} & a_{24} & & \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & \\ & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ & & a_{53} & a_{54} & a_{55} & a_{56} \\ & & & a_{64} & a_{65} & a_{66} \end{pmatrix}$$

$$L=\begin{pmatrix} \times & \times & a_{11} & a_{12} & a_{13} \\ \times & a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{53} & a_{54} & a_{55} & a_{56} & \times \\ a_{64} & a_{65} & a_{66} & \times & \times \end{pmatrix}$$

图6.12　带状阵及其压缩存贮

图中画“×”处的元素可以是任意数. 由于消元将在L区域上进行，但原矩阵中的第1列的元素在L中是被排在第$m+1$条上行对角线上的，因此在第一次消元之前，应先在前m行中，分别将第$1,2,\cdots,m$行左移$m,m-1,\cdots,1$次，使L的前m行还原成矩阵原来的形状，其后m行中原来画“×”处的元素亦应同时清零，即使L在第一次消元以前的初态为

$$L=\begin{pmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} & 0 \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{53} & a_{54} & a_{55} & a_{56} & 0 \\ a_{64} & a_{65} & a_{66} & 0 & 0 \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\}(m+1) \\ \\ \\ \\ \end{matrix}.$$

当在L的前$m+1$行的第1列中进行列主元消去法时，主

元的选取、换行以及实现消元的步骤与通常的列主元消去法步骤相同。第1次消元结束时，L的前$m+1$行的形状及数值意义为

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1m+1}^{(1)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2m+1}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & a_{m+1,2}^{(1)} & \cdots & a_{m+1,m+1}^{(1)} \end{pmatrix},$$

此时第1行中的信息即第一次消元结束时的输出，它们在回代时还要用到，应予保留，而当把第$2\sim m-1$行左移一列时，L在第$2\sim m+2$行上的元素正好合于第2次消元时的需要，之后，消元即应在第$2\sim m+2$行上进行，仿此直到消元过程进行完毕。当然，在行与行间进行换行及消元步骤时，对方程组的右端项向量b亦应进行相应的处理。第i次消元时，参加消元的行数一般应是$m+1$行，即第i行到第$i+m$行，当$i+m>n$时，则应是第i行到第n行，参加消元的列数可以一直是$2m+1$列，消元结束时，开始回代，而形成的解在b中，算法的具体步骤如下：

算法6.3.1 等带宽阵的列主元消去法：

1）置$m_1=m+1$，$m_2=2m+1$；

2）对$r=1,2,\cdots,m_1-1$，

　2.1）对$i=1,\cdots,m_1-r$，

　　2.1.1）置$L[r,j-1]=L[r,j]\ (j=1,\cdots,m_2)$，

　　2.1.2）置$L[r,m_2]=L[n+1-r,m_2-i+1]$
$=0$，

　　2.1.3）NEXT i，

　2.2）NEXT r；

3）对$i=1,\cdots,n-1$，

3.1) 置 $k=i$,

3.2) 对 $r=i+1, \cdots, m_1$,

3.3) 如果 $|L[r,1]|>|L[k,1]|$ 则置 $k=r$,

3.4) NEXT r;

4) 如果 $k \neq i$ 则转 5), 否则转 7);

5) $b[i] \longleftrightarrow b[k]$;

6) $L[i,j] \longleftrightarrow L[k,j] \quad j=1, \cdots, m_2$;

7) 置 $b[i]=b[i]/L[i,1]$;

置 $L[i,j]=L[i,j]/L[i,1]$, $(j=2, \cdots, m_2)$.

8) 对 $r=i+1, \cdots, m_1$,

8.1) 置 $t=L[r,1], b[r]=b[r]-t*b[i]$,

8.2) $L[r,j-1]=L[r,j]-t*L[i,j]$, $(j=2, \cdots, m_2)$,

8.3) 置 $L[r,m_2]=0$,

8.4) NEXT r;

9) 如果 $m_1 \neq n$ 则置 $m_1 \leftarrow m_1+1$;

10) NEXT i;

11) 置 $b[n]=b[n]/L[n,1]$, $j_m=2$;

12) 对 $r=n-1, n-2, \cdots, 1$,

12.1) $b[r]=b[r]-L[r,j]*b[r-1+j]$,

$(j=2, \cdots, j_m)$,

12.2) 如果 $j_m \neq m$ 则置 $j_m \leftarrow j_m+1$,

12.3) NEXT r.

3.2 对称等带宽矩阵的消去法

由于不带行交换的高斯消去法在整个消元过程中，每个行上第一个非零元左侧的元素将始终保持是零，因而当矩阵 A 是正定对称等带宽矩阵时，则在分解式

$$A=LDL^T \tag{6.3.3}$$

中，L将与A有同样的带宽，这里L是下三角阵，D是对角阵，它们都是非奇异的。当规定

$$d_{ii}=1/l_{ii} \tag{6.3.4}$$

时，分解式(6.3.3)是唯一的。

设m为A的半带宽，比较(6.3.3)两端的下三角部分，有

$$\begin{aligned} a_{ij} &= \sum_{k=1}^{n} l_{ik}d_{kk}l_{jk} \quad (j\leqslant i) \\ &= \sum_{k=t_i}^{j} l_{ik}l_{jk}/l_{kk}, \end{aligned}$$

得

$$\begin{aligned} l_{ii} &= a_{ii} - \sum_{k=t_i}^{i-1} l_{ik}^2/l_{kk} \\ l_{ri} &= a_{ri} - \sum_{k=t_r}^{i-1} l_{rk}l_{ik}/l_{kk} \qquad (6.3.5) \\ t_r &= \max(1, r-m). \end{aligned}$$

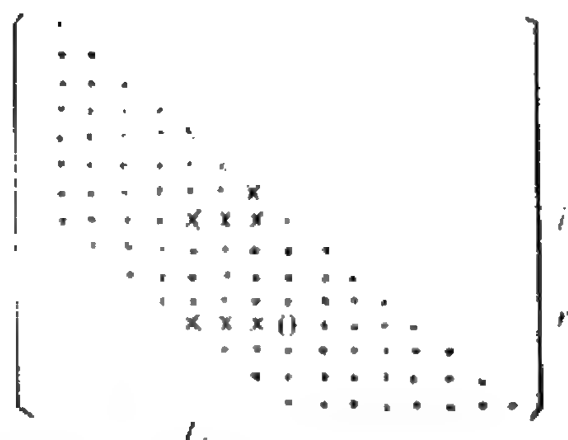


图6.13 计算l_{ri}所需的信息

由图6.13中所示的计算l_{ri}所需的信息，包括画“×”处的元素，这是L中在计算l_{ri}之前已经算出的信息；画“∘”处的元素是A阵中的a_{ri}，可见L中的元素可以逐列按由上而下的次序也可以逐行按自左至右的次序进行计算。当新的l_{ri}算出以后，即可用它来取代a_{ri}的存贮位置。由

$$LDL^T x = b, L^T x = \overline{b}, LD\overline{b} = b,$$

可得

$$\overline{b}_i = b_i - \sum_{k=m_i}^{i-1} l_{ik}\overline{b}_k / l_{kk}$$

$$x_i = \left(\overline{b}_i - \sum_{k=i}^{p_i-1} l_{ki} x_k\right) / l_{ii} \qquad (6.3.6)$$

$m_i = \max(1, i-m)$, $p_i = \min(n, i+m)$。

当A的下三角部分以压缩形式存于二维数组$L[1:n, 1:m+1]$时，其地址检索关系为：

$$A[i,j] = L[i', j'], \qquad (6.3.7)$$

$$i' = i, j' = j - i + m + 1.$$

于(6.3.5)、(6.3.6)中换a_{ri}, l_{ik}, …为$L[r, i-r+m+1]$、$L[i, k-i+m+1]$, …, 则得压缩存贮形式下的矩阵分解及求解的计算公式。

算法6.3.2 对称等带宽矩阵的消去法：

1) 置$m_1 = m+1$;

2) 对$i = 1, 2, \cdots, n$,

2.1) 置$t = \max(1, i-m)$,

2.2) 对$j = t, t+1, \cdots, i$,

2.2.1) $L[i, j-i+m_1] \leftarrow L[i, j-i+m_1]$

$$-\sum_{k=t}^{j-1} L[i, k-i+m_1] L[j, k-j+m_1] / L[k, m_1],$$

2.3) $b[i] \leftarrow \left(b_i - \sum_{k=t}^{i-1} L[i, k-i+m_1] b[k]\right) / L[i, m_1]$;

3) NEXT i;

4) 对$i = n, n-1, \cdots, 1$,

4.1) 置$t = \min(n, i+m)$,

4.2) $b[i]=\left(b[i]-\sum_{k=i+1}^{t}L[k,i-k+m_1]*b[k]\right)/L[i,m_1]$;

5) NEXT i。

3.3 对称变带宽矩阵的消去法

当A是对称正定的变带宽矩阵时，其分解式

$$A=LDL^T$$

中，L在每个行上的半带宽应与A在同行上的一致，这里L、D的意义与3.2段中的规定是一样的。

在A是变带宽的情形下，L的计算公式与A是等带宽情形下的区别仅在于：应把公式(6.3.5)中求和符号的下限$t_r=\max(1,r-m)$换为

$$t_r=\max(m_i,m_r)\ (i\leqslant r), \tag{6.3.8}$$

这里m_i,m_r分别是第i行第r行上第1个非零元的列号，它们可分别用第i，r行上的半带宽计算出来。于是，可写出相应于(6.3.5)的公式：

$$\begin{aligned}
&l_{ii}=a_{ii}-\sum_{k=t_i}^{i-1}l_{ik}^2/l_{kk},\\
&l_{ri}=a_{ri}-\sum_{k=t_r}^{i-1}l_{rk}l_{ik}/l_{kk},\\
&t_r=\max(m_i,m_r),\\
&m_i=i-\beta_i m_r=r-\beta_r,
\end{aligned} \tag{6.3.9}$$

和

$$\begin{aligned}
&\bar{b}_i=b_i-\sum_{k=m_r}^{i-1}l_{ik}\bar{b}_k/l_{kk},\\
&x_i=\left(\bar{b}_i-\sum_{k=i}^{n}l_{ki}x_k\right)/l_{ii}.
\end{aligned} \tag{6.3.10}$$

当变带宽矩阵的下三角部分以压缩形式存放在一维数组$L[1:S]$时，据地址检索关系：

$$\begin{aligned}
&L(s)=A[i,j],\\
&s=d[i]+j-i,
\end{aligned} \tag{6.3.11}$$

$$i=1,2,\cdots,n,\ j=m_{ij},\cdots,i.$$

这里

$$s=\sum_{i=1}^{n}(\beta_i+1),\tag{6.3.12}$$

而 $d(k)$ 是 A 的第 k 个对角元在 L 数组中的位置。对于(6.3.9)、(6.3.10)，分别将 $l_{ik},l_{kk},\cdots$ 换为 $L[d[i]+k-i]$, $L[d[k]]$，即可得出对称变带宽存贮下的矩阵分解及方程组求解的计算公式。注意到公式(6.3.10)中计算 x_i 时要用到 L 在第 i 列中的数据，而 L 区域的形状对列是非凸的，因此检索关系(6.3.11)只有当 (i,j) 属于变带宽区域时才能使用，否则 l_{ij} 便应是零。判别 (i,j) 是否属于变带宽区域只需看如下不等式

$$m_i=i-d[i]+d[i-1]+1\leqslant j\leqslant i\tag{6.3.13}$$

是否成立即可。

算法6.3.3 变带宽对称矩阵的消去法：

1) 对 $i=1,2,\cdots,n$,

 1.1) 置 $i_0=d[i]-i$, $m_i=i-(d[i]-d[i-1])+1$,

 1.2) 对 $j=m_i,\cdots,i$,

  1.2.1) 置 $j_0=d[j]-j$, $m_j=j-(d[j]-d[j-1])+1$, $m_{ij}=\max(m_i,m_j)$, $ij=i_0+j$,

  1.2.2) $L[ij]\leftarrow L[ij]-\sum\limits_{k=m_{ij}}^{j-1}L[i_0+k]L[j_0+k]/L[d[k]]$,

  1.2.3) NEXT j;

 1.3) $b[i]\leftarrow\left(b[i]-\sum\limits_{k=m_i}^{i-1}L[i_0+k]b[k]\right)/L[d[k]]$,

 1.4) NEXT i;

2) $b[n]\leftarrow b[n]/L[d[n]]$;

3) 对 $i=n-1,\cdots,1$,

 3.1) 置 $s=0$,

3.2) 对 $k=i+1, \cdots, n$,

3.2.1) 置 $k_0=d[k]-k$,

$m_k=k-(d[k]-d[k-1])+1$,

3.2.2) 如果 $m_k \leqslant i \leqslant k$，则

$s \leftarrow s+L[k_0+i]b[k]$,

3.2.3) NEXT k,

3.3) $b[i] \leftarrow (b[i]-S)/L[d[i]]$;

4) NEXT i.

第六章习题

6.1 当按压缩存贮方式把对称正定矩阵存入内存时，试

1) 建立压缩存贮矩阵与原矩阵的地址检索关系；

2) 导出用压缩存贮矩阵元素表示的有关消元及回代过程公式.

6.2 证明，在对称正定带状阵的 Cholesky 分解

$$A=LL^T$$

中，L的任意一行上的半带宽与A的同一行上的半带宽相同.

6.3 在链式存贮中，如果因行列信息出错不能检索出元素的地址，试编出能诊断这一情况并予以警告的程序.

6.4 在链式存贮中，设欲在第j列增加的元素是本列的第一个非零元或最末一个非零元，试写出相应的检索步骤.

6.5 在链式存贮中，设第j列的某元素的行号为i，应依怎样的步骤找出同列中离它最近的非零元信息条首址?

6.6 设图 6.14 为某一对称正定矩阵的无向图为

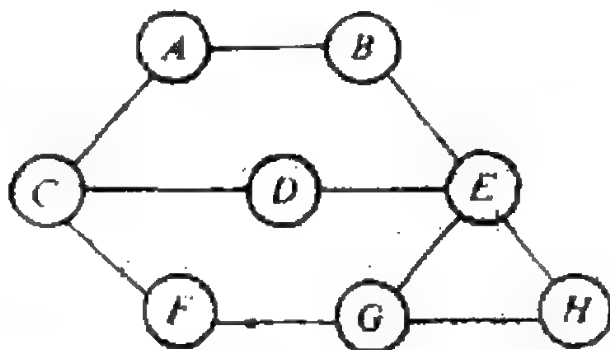


图 6.14

试分别以 $x_C, x_D, x_G, x_B, x_H, x_F, x_A, x_F$ 及 $x_H, x_F, x_D, x_C, x_E, x_B, x_A$ 为次序写出相应矩阵的Clolesky 分解的下三角阵中的非零元分布的情况，两种排序中以何者为好？能否据此提出一般排序的建议？

6.7　用 Tewarsen 方法编制第 k 次消元过程中的求列主元的子程序。

6.8　用 Markowitz 方法编制第 k 次消元过程中的求列主元的子程序。

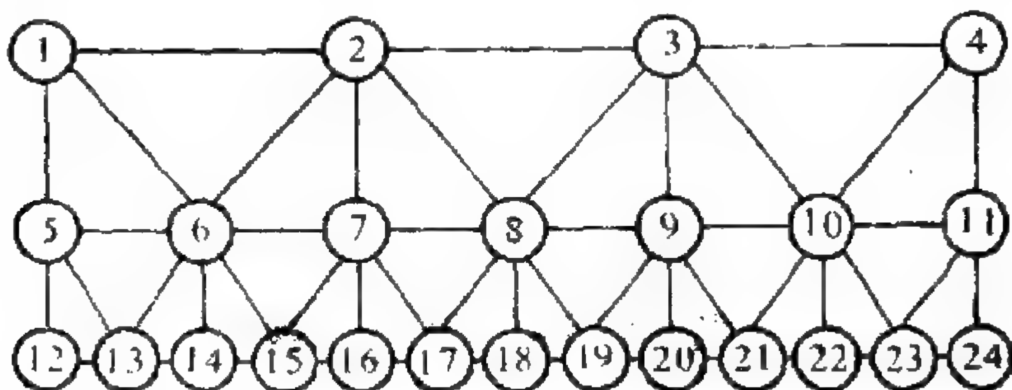


图6.15

6.9　按图右中节点编号的次序写出相应的对称正定矩阵各行上的半带宽并求其外形。

6.10　据上图以节点 12为根写出分层结构中各层的节点，按 CM 方法和 RCM 方法重新排序，求出两种排序下的矩阵的带宽和外形。

参　考　书

1. R. P. Tewarson,"*Sparse Matrices*", Academic Press, New York (1973).
2. J. K. Reid,"*Sparse Matrices*" 引自"*The State of the art in numerical Analysis*", Academic Press, London, New York San Francis co.(1977).
3. A. Jennings, "*Matrix Computation for Engineers and Scientists*",John wiley & Sons, London,New york,Sydney,Toronto (1977).
4. *Iain S. Duff* "稀疏矩阵研究的综述"，应用数学与计算数学 1980 年第 1,2 期，唐路新译自 Proc. IEEE, 1977, 65, No. 4.
5. "数值代数的一些发展概况" 中国科学院计算中心图书情报室 (1978).

第七章　线性方程组的迭代解法

§1　引　　言

线性方程组的直接解法，对于低阶方程组是很有成效的。对于高阶稀疏方程组，目前虽然也介绍了各种有效的方法，但是由于有存贮量大，编程序的技巧要求较高等缺陷，在实践中不免会遇到一些困难。迭代法则是克服这方面缺点的有效方法。迭代法与直接法的根本区别在于，后者即使每一步的计算都是精确的运算，得到的仍是近似解，而前者则是从一系列的近似解中去获得满足精度要求的近似解的。由于迭代法必须考虑收敛性和误差，所以需要根据问题的背景和条件去选择合适的方法，即收敛速度较快的方法。这是研究迭代法的两个重要问题。

现在我们以三阶方程组为例，对数值分析中介绍过的迭代方法进行回顾。

设三阶方程组为

$$\begin{cases} a_{11}x_1+a_{12}x_2+a_{13}x_3=b_1, \\ a_{21}x_1+a_{22}x_2+a_{23}x_3=b_2, \\ a_{31}x_1+a_{32}x_2+a_{33}x_3=b_3. \end{cases} \tag{7.1.1}$$

设$a_{ii}\neq 0\,(i=1,\ 2,\ 3)$，将(7.1.1)变形为

$$\begin{cases} x_1=-\dfrac{a_{12}}{a_{11}}x_2-\dfrac{a_{13}}{a_{11}}x_3+\dfrac{b_1}{a_{11}}, \\ x_2=-\dfrac{a_{21}}{a_{22}}x_1-\dfrac{a_{23}}{a_{22}}x_3+\dfrac{b_2}{a_{22}}, \\ x_3=-\dfrac{a_{31}}{a_{33}}x_1-\dfrac{a_{32}}{a_{33}}x_2+\dfrac{b_3}{a_{33}}. \end{cases} \tag{7.1.2}$$

把它写成矩阵形式，则为

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}^{-1} & & \\ & a_{22}^{-1} & \\ & & a_{33}^{-1} \end{pmatrix} \begin{pmatrix} 0 & -a_{12} & -a_{13} \\ -a_{21} & 0 & -a_{23} \\ -a_{31} & -a_{32} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$+ \begin{pmatrix} a_{11}^{-1} & & \\ & a_{22}^{-1} & \\ & & a_{3}^{-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

将(7.1.2)写成迭代格式，就得到简单迭代法的迭代格式：

$$\begin{cases} x_1{}^{(m)} = -\dfrac{a_{12}}{a_{11}} x_2{}^{(m-1)} - \dfrac{a_{13}}{a_{11}} x_3{}^{(m-1)} + \dfrac{b_1}{a_{11}}, \\ x_2{}^{(m)} = -\dfrac{a_{21}}{a_{22}} x_1^{(m-1)} - \dfrac{a_{23}}{a_{22}} x_3^{(m-1)} + \dfrac{b_2}{a_{22}}, \\ x_3^{(m)} = -\dfrac{a_{31}}{a_{33}} x_1^{(m-1)} - \dfrac{a_{32}}{a_{33}} x_2^{(m-1)} + \dfrac{b_3}{a_{33}}. \end{cases} \qquad (7.1.3)$$

对于迭代格式(7.1.3)，还可以进行改进。由于在做(7.1.3)的第2式时，$x_1^{(m)}$已经计算出来了，在做第3式时，$x_2^{(m)}$也计算出来了，因此，我们可以将(7.1.3)改写成

$$\begin{cases} x_1^{(m)} = -\dfrac{a_{12}}{a_{11}} x_2^{(m-1)} - \dfrac{a_{13}}{a_{11}} x_3^{(m-1)} + \dfrac{b_1}{a_{11}}, \\ x_2^{(m)} = -\dfrac{a_{21}}{a_{22}} x_1^{(m)} - \dfrac{a_{23}}{a_{22}} x_3^{(m-1)} + \dfrac{b_2}{a_{22}}, \\ x_3^{(m)} = -\dfrac{a_{31}}{a_{33}} x_1^{(m)} - \dfrac{a_{32}}{a_{33}} x_2^{(m)} + \dfrac{b_3}{a_{33}}. \end{cases} \qquad (7.1.4)$$

这就是 GS 法的迭代格式。如果在这个迭代格式中引进一个加速迭代参数，即所谓**松弛因子**，并将(7.1.4)改写成

$$
\begin{cases}
x_1^{(m)} = \omega\left(-\dfrac{a_{12}}{a_{11}}x_2^{(m-1)} - \dfrac{a_{13}}{a_{11}}x_3^{(m-1)} + \dfrac{b_1}{a_{11}}\right) + (1-\omega)x_1^{(m-1)}, \\
x_2^{(m)} = \omega\left(-\dfrac{a_{21}}{a_{22}}x_1^{(m)} - \dfrac{a_{23}}{a_{22}}x_3^{(m-1)} + \dfrac{b_2}{a_{22}}\right) + (1-\omega)x_2^{(m-1)}, \\
x_3^{(m)} = \omega\left(-\dfrac{a_{31}}{a_{33}}x_1^{(m)} - \dfrac{a_{32}}{a_{33}}x_2^{(m)} + \dfrac{b_3}{a_{33}}\right) + (1-\omega)x_3^{(m-1)}.
\end{cases}
\tag{7.1.5}
$$

这就是**逐次松弛法**的迭代格式。可简记成SOR法。如果设

$$
D = \begin{pmatrix} a_{11} & & \\ & a_{22} & \\ & & a_{33} \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 0 & 0 \\ -a_{21} & 0 & 0 \\ -a_{31} & -a_{32} & 0 \end{pmatrix},
$$

$$
U = \begin{pmatrix} 0 & -a_{12} & -a_{13} \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix},
$$

则简单迭代法的迭代格式为

$$
\boldsymbol{x}^{(m)} = D^{-1}(L+U)\boldsymbol{x}^{(m-1)} + D^{-1}\boldsymbol{b}, \tag{7.1.6}
$$

GS法的迭代格式为

$$
\boldsymbol{x}^{(m)} = D^{-1}\{L\boldsymbol{x}^{(m)} + U\boldsymbol{x}^{(m-1)}\} + D^{-1}\boldsymbol{b},
$$

或

$$
(I - D^{-1}L)\boldsymbol{x}^{(m)} = D^{-1}U\boldsymbol{x}^{(m-1)} + D^{-1}\boldsymbol{b}
$$

即

$$
\boldsymbol{x}^{(m)} = (I - D^{-1}L)^{-1}D^{-1}U\boldsymbol{x}^{(m-1)} + (I - D^{-1}L)^{-1}D^{-1}\boldsymbol{b}; \tag{7.1.7}
$$

SOR 法的迭代格式为

$$
\boldsymbol{x}^{(m)} = \omega(D^{-1}L\boldsymbol{x}^{(m)} + D^{-1}U\boldsymbol{x}^{(m-1)} + D^{-1}\boldsymbol{b}) + (1-\omega)\boldsymbol{x}^{(m-1)},
$$

或

$$
(I - \omega D^{-1}L)\boldsymbol{x}^{(m)} = \omega D^{-1}U\boldsymbol{x}^{(m-1)} + (1-\omega)\boldsymbol{x}^{(m-1)} + \omega D^{-1}\boldsymbol{b}
$$

即

$$
\boldsymbol{x}^{(m)} = (I - \omega D^{-1}L)^{-1}\{(1-\omega)I + \omega D^{-1}U\}\boldsymbol{x}^{(m-1)} + (I - \omega D^{-1}L)^{-1}\omega D^{-1}\boldsymbol{b}. \tag{7.1.8}
$$

当$\omega = 1$时，SOR法化为GS法。

可以把三种迭代格式写成统一的形式：

$$\boldsymbol{x}^{(m)} = B\boldsymbol{x}^{(m-1)} + \boldsymbol{g}, \qquad (7.1.9)$$

其中矩阵B称为迭代法(7.1.9)中的**迭代矩阵**。

对于n阶方程组

$$A\boldsymbol{x} = \boldsymbol{b}, \qquad (7.1.10)$$

其中A是n阶方阵，$\boldsymbol{x}$，$\boldsymbol{b}$为n维列向量，如果令$A = D - L - U$，其中

$$D = \begin{pmatrix} a_{11} & & & 0 \\ & a_{22} & & \\ 0 & & \ddots & \\ & & & a_{nn} \end{pmatrix},$$

$$L = \begin{pmatrix} 0 & & & 0 \\ -a_{21} & 0 & & \\ \vdots & \ddots & \ddots & \\ -a_{n1} & \cdots & -a_{n\,n-1} & 0 \end{pmatrix},$$

$$U = \begin{pmatrix} 0 & -a_{12} & \cdots & -a_{1n} \\ & 0 & \ddots & \vdots \\ & & 0 \ \ddots & -a_{n-1\,n} \\ & & & 0 \end{pmatrix},$$

则迭代法(7.1.6)、(7.1.7)、(7.1.8)、(7.1.9)同样成立。

对于任意的初始值$\boldsymbol{x}^{(0)}$，由(7.1.9)得出的向量序列$\boldsymbol{x}^{(1)}$，$\boldsymbol{x}^{(2)}$，…，$\boldsymbol{x}^{(m)}$，…，如果收敛于某个向量$\boldsymbol{x}^*$，则$\boldsymbol{x}^*$就是方程组(7.1.1)的精确解。因此，迭代法实际上是一种从已有近似解去计算新近似解的法则。从(7.1.9)看出，新近似解$\boldsymbol{x}^{(m)}$是已

有近似解 $x^{(m-1)}$ 的线性函数，这种迭代法叫**线性迭代法**。在本章的最后一节我们还将介绍**非线性迭代法**.此外，由于(7.1.9)中 $x^{(m)}$ 仅是前一次近似解 $x^{(m-1)}$ 的函数，这种迭代法又叫做**一阶迭代法**。如果不仅依赖前一次的近似解，而且还依赖前面相邻 K 次的近似解 $x^{(m-1)}$，$x^{(m-2)}$，$\cdots$，$x^{(m-k)}$，则称为**K 阶迭代法**。例如Chebysher半迭代法就是二阶迭代法。从(7.1.9)中还可看出，无论 m 为何值，计算 $x^{(m)}$ 的规则均是不变的，这种迭代法称为**定常迭代法**。Chebyshev 半迭代法还是一种非定常的线性迭代法。

我们构造迭代格式的目的是要求原方程组的解。因此，在构造迭代格式时，必须注意迭代法的极限是否是原方程组的解。如果对于任取的初始向量 $x^{(0)}$，由(7.1.9)所产生的向量序列 $\{x^{(m)}\}$ 都有相同的极限，并且其极限就是原方程组的解 x^*，则称迭代格式(7.1.9)是**收敛的**，否则称为**不收敛的**。

定理7.1.1 对于任意右端向量 g 和初始向量 $x^{(0)}$，迭代法(7.1.9) 收敛的充要条件是 $S(B)<1$。这里 $S(B)$ 表示 B 的谱半径。

证明 必要性。设迭代法(7.1.9)收敛且极限是 x^*，即

$$x^*=Bx^*+g.$$

将上式与(7.1.9)相减得到

$$x^{(m)}-x^*=B(x^{(m-1)}-x^*).$$

令

$$r^{(m)}=x^{(m)}-x^*, \tag{7.1.11}$$

则有

$$r^{(m)}=Br^{(m-1)}=B^2r^{(m-2)}=\cdots=B^mr^{(0)}. \tag{7.1.12}$$

据假设

$$\lim_{m\to\infty}r^{(m)}=\lim_{m\to\infty}B^mr^{(0)}=0,$$

故必有 $\lim\limits_{m\to\infty}B^m=0$，因此 $S(B)<1$，

充分性。设 $S(B)<1$，则 $I-B$ 非奇异，从而方程组

$$(I-B)x=g$$

有唯一解，记为 x^*。由 $S(B)<1$，可得 $\lim\limits_{m\to\infty}B^m=0$，故有

$$\lim_{m\to\infty}r^{(m)}=\lim_{m\to\infty}B^m r^{(0)}=0,$$

即

$$\lim_{m\to\infty}x^{(m)}=x^*.$$

对于迭代法，不但要知道它是否收敛，而且还必须研究它的收敛速度。

由(7.1.12)，可得

$$\|r^{(m)}\|\leqslant\|B^m\|\ \|r^{(0)}\|,$$

因此，$\|B^m\|$ 的大小可以决定 $r^{(m)}$ 收敛于零向量的速度。如果要使 $\|r^{(m)}\|$ 比 $\|r^{(0)}\|$ 减小 ε 倍 $(0<\varepsilon<1)$

即

$$\|r^{(m)}\|\leqslant\varepsilon\|r^{(0)}\|,$$

为此仅要求

$$\|B^m\|\leqslant\varepsilon\text{或}(\|B^m\|^{\frac{1}{m}})^m\leqslant\varepsilon.$$

两边取对数，得到

$$\ln\varepsilon\geqslant m\ln(\|B^m\|^{\frac{1}{m}}),$$

即

$$m\geqslant-\ln\varepsilon/(-\frac{1}{m}\ln\|B^m\|).\tag{7.1.13}$$

由此可见，要达到精度为 ε 的要求，所需迭代的最少次数与 $(-\ln\|B^m\|/m)$ 成反比，这个量愈大，所需的迭代次数就愈小。因此，我们可以用这个量来衡量迭代格式的收敛速度。

定义7.1.1 称 $R_m(B)=-\dfrac{1}{m}\ln\|B^m\|$ 为迭代法(7.1.9)的**平均收敛速度**。

由于平均收敛速度难于求出，因此我们往往用一粗略的估计式来代替它。若 B 仅是有线性初等因子的矩阵，则必存在 B 的线性无关的特征向量系 x_1，$x_2,\cdots,x_n$ 和相应的特征值 λ_1,λ_2,

$\cdots$，λ_n。对于任意初始误差向量$r^{(0)}$，设

$$r^{(0)}=\sum_{i=1}^{n}\alpha_i x_i,$$

代入(7.1.12)得到

$$r^{(m)}=\sum_{i=1}^{n}\alpha_i B^m x_i=\sum_{i=1}^{n}\alpha_i\lambda_i^m x_i$$

$$=(S(B))^m\sum_{i=1}^{n}\left(\frac{\lambda_i}{S(B)}\right)^m\alpha_i x_i,$$

两边取范数，便有

$$\|r^{(m)}\|\leqslant(S(B))^m\left\|\sum_{i=1}^{n}\left(\frac{\lambda_i}{S(B)}\right)^m\alpha_i x_i\right\|.$$

当m很大时，$\|r^{(m)}\|$收敛于零的速度取决于$(S(B))^m$的大小，如果要求$\|r^{(m)}\|\leqslant\varepsilon$，那么，迭代的次数$m$必须近似地满足

$$(S(B))^m\leqslant\varepsilon,$$

即 $$m\geqslant(-\ln\varepsilon)/(-\ln(S(B))),\qquad(7.1.14)$$

因而所需要的迭代次数m与$(-\ln(S(B)))$成反比。同时，还可以证明$\lim\limits_{m\to\infty}(\|B^m\|)^{\frac{1}{m}}=S(B)$。

定义7.1.2 称$R(B)=-\ln(S(B))$为迭代法(7.1.9)的**渐近收敛速度**。简称**收敛速度**。

利用$R(B)$可以比较各种迭代法收敛的快慢程度。显然，谱半径$S(B)$愈小，则收敛速度将愈快。但是迭代次数的比较可靠的估计还是(7.1.13)，而(7.1.14)仅是粗略的估计式。特别，当B的按模最大特征值所对应的Jordan块的阶数大于1时，实际迭代的次数往往要比估计的次数大得多。应用(7.1.14)时必须注意这一点〔参看**Iteratire solution of large linear systems** P85—P87。〕

§2 逐次松弛法

逐次松弛法是一阶线性定常迭代法，对于一类特殊的矩阵，如果松弛因子ω选得恰当，它的收敛速度会比Jacobi方法和GS方法快得多。通常我们把$\omega>1$的松弛迭代法称为**超松弛迭代法**，而把$0<\omega<1$的松弛迭代法称为**低松弛迭代法**。不加区别，则统称为**逐次松弛法**。松弛因子选在什么范围内松弛法才能收敛？是我们最关心的问题之一。下面我们将在数值分析的基础上，较详细地来讨论逐次松弛法的收敛性和最佳松弛因子的选取问题。

2.1 逐次松弛法的收敛性

定理7.2.1 当ω取实数时，逐次松弛法(7.1.8)收敛的必要条件是$0<\omega<2$。 (7.2.1)

证明 设(7.1.8)的迭代矩阵为

$$J_\omega=(I-\omega D^{-1}L)^{-1}[(1-\omega)I+\omega D^{-1}U], \qquad (7.2.2)$$

则

$$\det(J_\omega)=\det((I-\omega D^{-1}L)^{-1})\det((1-\omega)I+\omega D^{-1}U)$$
$$=(1-\omega)^n.$$

因此，矩阵J_ω的所有特征值之积为$(1-\omega)^n$，故

$$S(J_\omega)\geqslant|1-\omega|. \qquad (7.2.3)$$

据定理7.1.1，SOR法收敛的充要条件为$S(J_\omega)<1$知，必须$|1-\omega|<1$，即$0<\omega<2$。

实用中碰到的系数矩阵，多数是对称正定的。例如，用有限元素法求解弹性结构的静力平衡问题，就可归结为解一个系数矩阵为对称正定的线性代数方程组的问题。对于这类系数矩阵，收敛性问题现已得到了较好的解决。

定理7.2.2 设A是实对称正定矩阵，则$0<\omega<2$时，逐次松弛法是收敛的。

证明 因为A是实对称正定阵，所以有

$$A=D-L-L^T, \tag{7.2.4}$$

$$J_\omega=(I-\omega D^{-1}L)^{-1}[(1-\omega)I+\omega D^{-1}L^T]. \tag{7.2.5}$$

设λ为J_ω的任一特征值，x为相应的特征向量，则

$$J_\omega x=\lambda x,$$

即

$$[(1-\omega)I+\omega D^{-1}L^T]x=\lambda(I-\omega D^{-1}L)x,$$

$$(1-\omega)x+\omega D^{-1}L^Tx=\lambda(x-\omega D^{-1}Lx).$$

两边用x^HD左乘，得

$$(1-\omega)x^HDx+\omega x^HL^Tx=\lambda(x^HDx-\omega x^HLx). \tag{7.2.6}$$

由于D为对角矩阵且对角线元素均为正实数，因此若令$p=x^HDx$，则p为正实数，又$(x^HL^Tx)^H=x^HLx$，所以若令$x^HLx=\alpha+i\beta$，则$x^HL^Tx=\alpha-i\beta$。代入(7.2.6)，得到

$$(1-\omega)p+\omega(\alpha-i\beta)=\lambda(p-\omega(\alpha+i\beta)).$$

因此

$$\lambda=\frac{(1-\omega)p+\omega\alpha-i\omega\beta}{p-\omega\alpha-i\omega\beta},$$

$$|\lambda|=\frac{[p-\omega(p-\alpha)]^2+\omega^2\beta^2}{(p-\omega\alpha)^2+\omega^2\beta^2},$$

由于

$$(p-\omega(p-\alpha))^2-(p-\omega\alpha)^2=p\omega(2-\omega)(2\alpha-p),$$

利用A的正定性可得

$$x^HAx=x^HDx-x^HLx-x^HL^Tx=p-2\alpha>0.$$

从而，当$0<\omega<2$时，便有

$$(p-\omega(p-\alpha))^2-(p-\omega\alpha)^2<0,$$

得

$$|\lambda|^2=\frac{(p-\omega(p-\alpha))^2+\omega^2\beta^2}{(p-\omega\alpha)^2+\omega^2\beta^2}<1,$$

所以

$$S(J_\omega)<1.$$

据定理 7.1.1，SOR 法收敛。

定义7.2.1 设A为n阶方阵，如对所有的$i(1\leqslant i\leqslant n)$均有

$$|a_{ii}|\geqslant\sum_{\substack{j=1\\j\neq i}}^{n}|a_{ij}| \tag{7.2.7}$$

成立，则称A为**对角占优阵**。如果至少有一个i值，(7.2.7) 有严格的不等号成立。则称A为**弱角对角占优阵**。若对所有的i值，(7.2.7) 均有严格不等号成立，则称A为**严格对角占优阵。**

定义7.2.2 如果存在n阶排列矩阵P，能使n阶方阵A化为：

$$PAP^T=\begin{pmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{pmatrix}, \tag{7.2.8}$$

其中A_{11}、A_{22}均为方阵，则称A为**可约矩阵**，否则称为**不可约矩阵。**

例如，对于矩阵

$$A=\begin{pmatrix}4 & 0 & -1 & 0\\ -1 & 3 & -1 & -1\\ -1 & 0 & 2 & 0\\ -1 & -1 & -1 & 1\end{pmatrix},$$

存在矩阵

$$P=\begin{pmatrix}0 & 0 & 0 & 1\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 1 & 0 & 0 & 0\end{pmatrix},$$

使得

$$PAP^T=\begin{pmatrix}1 & -1 & -1 & -1\\ -1 & 3 & -1 & -1\\ 0 & 0 & 2 & -1\\ 0 & 0 & -1 & 4\end{pmatrix},$$

故A为可约矩阵。

如果线性方程组$Ax=b$的系数矩阵A为可约矩阵，那么，它可改写为

$$PAP^T(Px)=Pb, \tag{7.2.9}$$

若分别把Px和Pb记为y和h，并把PAP^T按(7.2.8)的形式进行分块，则(7.2.9)可以写成为

$$\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}=\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \tag{7.2.10}$$

即系数矩阵为可约矩阵的方程组，可以约化为两个较低阶的方程组来求解。又如

$$A=\begin{pmatrix} 4 & -1 & 0 & 0 \\ -1 & 3 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

为不可约的弱对角占优阵，其$\det(A)=8\neq 0$。不可约的弱对角占优阵是否一定为非奇异矩阵?可以证明如下定理(留作习题)。

定理7.2.3 若A为不可约的弱对角占优阵，则A必为非奇异阵，即$\det(A)\neq 0$。

定理7.2.4 当系数矩阵A为不可约弱对角占优矩阵且松弛因子ω满足$0<\omega\leqslant 1$，则SOR法必定收敛。

证明 设SOR法的迭代矩阵为

$$J_\omega=(I-\omega D^{-1}L)^{-1}((1-\omega)I+\omega D^{-1}U), \tag{7.2.11}$$

其中$A=D-L-U$须证$S(J_\omega)<1$。

用反证法，设J_ω有一个特征值λ满足$|\lambda|\geqslant 1$，则$J_\omega-\lambda I$应为奇异矩阵，即$\det(J_\omega-\lambda I)=0$，并且

$$\begin{aligned} J_\omega-\lambda I &= (I-\omega D^{-1}L)^{-1}[(1-\omega)I+\omega D^{-1}U]-\lambda I \\ &= (I-\omega D^{-1}L)^{-1}[(1-\omega)I+\omega D^{-1}U-\lambda(I-\omega D^{-1}L)] \end{aligned}$$

$$= (I-\omega D^{-1}L)^{-1}[(1-\omega-\lambda)I+\omega D^{-1}U+\lambda\omega D^{-1}L]$$

$$= (1-\omega-\lambda)(I-\omega D^{-1}L)^{-1}\left[I-\frac{\omega}{\omega+\lambda-1}D^{-1}U-\frac{\lambda\omega}{\omega+\lambda-1}D^{-1}L\right], \tag{7.2.12}$$

其中$1-\omega-\lambda\neq 0$和$(I-\omega D^{-1}L)^{-1}$为非奇异矩阵。由于$A=D-L-U$为不可约弱对角占优阵，所以，$D^{-1}A=I-D^{-1}L-D^{-1}U$也为不加约弱对角占优，又由于矩阵$D^{-1}A$中的零元素及其位置与矩阵$I-\frac{\omega}{\lambda+\omega-1}D^{-1}U-\frac{\lambda\omega}{\lambda+\omega-1}D^{-1}L$中的零元素及其位置一样，故后者也是不可约矩阵．此外又因为

$$\lambda=(\lambda+\omega-1)+(1-\omega),$$

$$|\lambda|\leqslant|\lambda+\omega-1|+(1-\omega),$$

$$|\lambda+\omega-1|\geqslant|\lambda|-(1-\omega),$$

所以

$$\left|\frac{\omega}{\lambda+\omega-1}\right|\leqslant\left|\frac{\lambda\omega}{\lambda+\omega-1}\right|\leqslant\frac{|\lambda\omega|}{|\lambda|-(1-\omega)}$$

$$\leqslant\frac{|\lambda|\omega}{|\lambda|-(1-\omega)|\lambda|}=1,$$

于是

$$I-\frac{\omega}{\lambda+\omega-1}D^{-1}U-\frac{\lambda\omega}{\lambda+\omega-1}D^{-1}L \tag{7.2.13}$$

也是弱对角占优阵。据定理7.2.3，(7.2.13)为非奇异矩阵，故(7.2.12)的右端为非奇异矩阵，这与左端为奇异矩阵相矛盾，于是$S(J_\omega)<1$。

由于当$\omega=1$时，SOR法即为GS法，因此有下述推论．

推论　当系数矩阵A为不可约对角占优时，GS方法收敛。

SOR法的收敛问题与松弛因子ω的选择密切相关．松弛因子选择得当，可以大大提高收敛速度。由于松弛法的应用比较广泛，所以如何挑选最佳的松弛因子，具有特别重要的意义。

2.2　最佳松弛因子的选取

前面我们已经证明，线性迭代法收敛的充要条件是，迭代矩阵的谱半径小于1，谱半径愈小，收敛就愈快。SOR 法的迭代矩阵的谱半径为$S(J_\omega)$，它是ω的函数。如何选取最佳松弛因子ω_{opt}，使得

$$S(J_{\omega_{opt}}) = \min_{\omega} S(J_\omega)$$

是我们需要研究的课题。但是，对于一般的系数矩阵，目前尚无确定ω_{opt}的理论结果。实际计算时，大多是用计算经验来试算确定的。仅对某些特殊类型的矩阵，才有确定ω_{opt}的理论公式。在此，我们重点讨论具有所谓"性质A"的矩阵的最佳松弛因子的选取问题。最后，将介绍用试算法确定最佳松弛因子的方法。

定义7.2.3 若能将n个自然数的集合$W=\{1,2,\cdots,n\}$，分成两个互不相交的子集S_1和S_2（即$S_1\cup S_2=W$，$S_1\cap S_2=\emptyset$），使得n阶矩阵A的所有非零非对角线元素a_{ij}（$a_{ij}\neq 0$，$i\neq j$）的足标对(i, j)，满足条件：$i\in S_1$，$j\in S_2$或$j\in S_1$，$i\in S_2$，则称此矩阵具有"性质A"。

例如，将线性方程组

$$\begin{pmatrix} 1 & -0.30009 & 0 & -0.30898 \\ -0.30009 & 1 & -0.46691 & 0 \\ 0 & -0.46691 & 1 & -0.27471 \\ -0.30898 & 0 & -0.27471 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5.32088 \\ 6.07624 \\ -8.80455 \\ 2.67600 \end{pmatrix} \tag{7.2.14}$$

的系数矩阵记为A，取$S_1=\{1, 3\}$，$S_2=\{2, 4\}$，容易验证矩阵A具有"性质A"。

如果把方程组(7.2.14)的第2，3两行互相交换并把2，3两列也互相交换，则得

$$\begin{pmatrix} 1 & 0 & -0.30009 & -0.30898 \\ 0 & 1 & -0.46691 & -0.27471 \\ -0.30009 & -0.46691 & 1 & 0 \\ -0.30898 & -0.27471 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_2 \\ x_4 \end{pmatrix}$$

$$= \begin{pmatrix} 5.32008 \\ -8.80455 \\ 6.07624 \\ 2.67600 \end{pmatrix}.$$

若令 $S_1 = \{1, 2\}$，$S_2 = \{3, 4\}$，仍可验证，新系数矩阵也具有性质 A。

定理 7.2.5 矩阵 A 具有“性质 A”的充要条件是存在排列矩阵 P，使得

$$PAP^T = \begin{pmatrix} D_1 & R \\ S & D_2 \end{pmatrix}, \tag{7.2.15}$$

其中 D_1，D_2 是对角矩阵。

证明 必要性。只需将具有“性质 A”的矩阵 A 的号码属于 S_1 的各行排在前面，把号码属于 S_2 的各行排在后面，并把号码属于 S_1 的各列也排在前面，把号码属于 S_2 的各列排在后面，便得(7.2.15)型的矩阵。

充分性。其证明作为习题。

对于线性方程组中的方程式和未知数，在次序进行重排以后，虽然方程组的精确解并不改变。但是，如果使用迭代法求解，其收敛性和收敛速度却要引起变化。同样对于具有性质 A 的矩阵，次序经过相应的变化后，虽然还保持“性质 A”，但其解的收敛性和收敛速度也要变化。那末，方程组和未知数的

次序应以怎样的排列为好呢？下面我们先看两个例子。

例7.2.1 对方程组(7.2.14)使用SOR法求解，取松弛因子$\omega=1.16$，实际迭代次数和结果见下表。

k	x_1	x_2	x_3	x_4	e_k
0	0	0	0	0	12.3095
1	6.1722	9.1970	-5.2320	3.6492	3.6659
2	9.6941	6.1177	-4.8999	4.4335	1.3313
3	8.3398	6.3188	-4.5914	3.9200	0.2302
4	8.4424	6.4480	-4.7151	4.0004	0.0767
5	8.5137	6.4202	-4.7068	4.0157	0.0288
6	8.4842	6.4253	-4.7005	4.0047	0.0051
7	8.4868	6.4288	-4.7031	4.0065	0.0016

例7.2.2 如果将方程组(7.2.14)的第3列与第4列交换，第3行与第4行交换，则得

$$
\begin{pmatrix}
1 & -3.30009 & -3.30898 & 0 \\
-3.30009 & 1 & 0 & -0.46691 \\
-3.30898 & 0 & 1 & -0.27471 \\
0 & -0.46691 & -0.27471 & 1
\end{pmatrix}
\begin{pmatrix} x_1 \\ x_2 \\ x_4 \\ x_3 \end{pmatrix}
=
\begin{pmatrix} 5.32088 \\ 6.07624 \\ 2.67600 \\ -8.80455 \end{pmatrix}. \tag{7.2.16}
$$

对这个方程组，用SOR方法求解，取松弛因子仍为$\omega=1.16$，其实际迭代次数及结果见下表。

K	x_1	x_2	x_3	x_4	e_i
0	0	0	0	0	12.3095
1	6.1722	-10.2133	3.6653	2.0618	6.2901
2	7.1996	-5.9369	5.7256	3.4629	1.9833
3	8.2639	-5.0442	6.2727	3.9047	0.4483
4	8.4330	-4.7645	6.3998	3.9837	0.0899
5	8.4785	-4.7153	6.4220	4.0031	0.0168
6	8.4859	-4.7050	6.4266	4.0059	0.0030
7	8.4874	-4.7032	6.4273	4.0065	0.0005

比较例7.2.1和例7.2.2不难看到，例7.2.2中 SOR 法收敛速度较快，这说明，线性方程组的方程和未知数的次序的排列不同，SOR法的收敛速度也将不同，那末，例 7.2.2 的排列次序有什么特点呢？为此，还将研究一些有关概念.

定义7.2.4 若能将n个自然数的集合 $W=\{1, 2, \cdots, n\}$ 分成t个互不相交的子集S_1，S_2，…，S_t(即 $W=\bigcup\limits_{k=1}^{t} S_k$，$S_i \cap S_j = 0$，$i \neq j$)，使得矩阵$A$的所有非零的非对角元 $a_{ij} \neq 0 \ (i \neq j)$，其足码对$(i,j)$，满足以下条件：若 $i \in S_k$，当$j<i$时，$j \in S_{k-1}$，而当 $j>i$ 时，$j \in S_{k+1}$，则称此矩阵A具有**相容次序**.

若取$S_1=\{1\}$，$S_2=\{2, 3\}$，$S_3=\{4\}$，容易验证(7.2.16)中的系数矩阵满足定义 7.2.4，即(7.2.16) 中的系数矩阵是具有相容次序的.方程组(7.2.16)的系数矩阵也具有“性质A”.

定理7.2.6 若矩阵具有相容次序，则必具有“性质A”.

证明 若矩阵具有相容次序，则存在满足定义 7.2.4 的子集S_1，S_2，…，S_t，令

$W_1 = \underset{i}{U}S_i$（i是奇数），$W_2 = \underset{i}{U}S_i$（i是偶数），则W_1与W_2就是定义7.2.3的两个子集S_1和S_2，即矩阵具有"性质A"。

例7.2.3 设矩阵A具有如下形式：

$$A = \begin{pmatrix} D_1 & U_1 & & \\ L_2 & D_2 & U_2 & \\ & \ddots & \ddots & U_{m-1} \\ & & L_m & D_m \end{pmatrix} \qquad (7.2.17)$$

其中D_i（$i=1, 2, \cdots, m$）为对角阵，L_{i+1}，U_i（$i=1, 2, \cdots, m-1$）为相应阶数的长方阵。若将所有对角线子块D_{2k}（$k=1, 2, \cdots$）的行编号记为S_1，将所有子块D_{2k-1}（$k=1, 2, \cdots$）的行编号记为S_2，则容易验证，集合S_1与S_2满足定义7.2.3，因此矩阵A具有"性质A"。如果将A中属于对角线子块D_k的行编号记为S_k，显然，$\bigcup_{i=1}^{m} S_k = W$，$S_i \cap S_j = 0$（$i \neq i$）。而且容易验证这样划分$W$的子集是满足定义7.2.4的条件的，所以矩阵(7.2.17)具有相容次序。

具有相容次序的矩阵A还有一些较重要的性质。为了证明这些性质，首先假设det（A）中不为零的一般项为

$$m(a) = \pm a_{1\alpha(1)} a_{2\alpha(2)} \cdots a_{n\alpha(n)}$$

对于$m(a)$中的每个元素$a_{i\alpha(i)}$，若$i > \alpha(i)$，则$a_{i\alpha(i)}$为主对角线下方的元素；如果$i < \alpha(i)$，则$a_{i\alpha(i)}$为主对角线上方的元素。

定理7.2.7 设A是具有相容次序的矩阵，l和j分别为$m(a)$中满足$i > \alpha(i)$和$i < \alpha(i)$的元素的个数，则$l = j$，即具有相容次序的矩阵A，其行列式det(A)中非零项中的元素，取在主对角线上方的个数与下方的个数是相等的。

证明 因A具有相容次序，存在满足定义7.2.4的子集S_1，

$S_2,\cdots,S_t$，即对于 $a_{i\alpha(i)}\neq 0$，若$i\in S_k$，则当$\alpha(i)>i$时，$\alpha(i)\in S_{k+1}$，而当$\alpha(i)<i$时，$\alpha(i)\in S_{k-1}$。现在定义一个n维向量：$\beta=(\beta_1,\ \beta_2,\ \cdots,\ \beta_n)^T$，其中$\beta_i$是这样确定的，当$i\in S_k$时，$\beta_i=k$。向量$\beta$具有如下性质：当$a_{i\alpha(i)}\neq 0$时，

$$\beta_i-\beta_{\alpha(i)}=\begin{cases}1, & i>\alpha(i);\\ -1, & i<\alpha(i).\end{cases}$$

于是 $$l=\sum_{\substack{i=1\\ i>\alpha(i)}}^{n}(\beta_i-\beta_{\alpha(i)}),\qquad j=\sum_{\substack{i=1\\ i<\alpha(i)}}^{n}(\beta_{\alpha(i)}-\beta_i).$$

因为 $$\begin{aligned} l-j&=\sum_{\substack{i=1\\ i>\alpha(i)}}^{n}(\beta_i-\beta_{\alpha(i)})-\sum_{\substack{i=1\\ i<\alpha(i)}}^{n}(\beta_{\alpha(i)}-\beta_i)\\ &=\sum_{i\neq\alpha(i)}(\beta_i-\beta_{\alpha(i)})+(\beta_i-\beta_i)\\ &=\sum_{i=1}^{n}(\beta_i-\beta_{\alpha(i)})=\sum_{i=1}^{n}\beta_i-\sum_{i=1}^{n}\beta_{\alpha(i)}=0,\end{aligned}$$

所以 $$l=j.$$

推论 设$A\in R^{n\times n}$具有相容次序且$D=\mathrm{diag}(a_{11},a_{22},\cdots,a_{nn})$，矩阵$E$为$A$的下三角部分的元素组成的下三角矩阵，矩阵$F$为$A$的上三角部分的元素组成的上三角矩阵，$\alpha$，$k$为任意非零实数，则

$$\det(\alpha E+\alpha^{-1}F+kD)\qquad(7.2.18)$$

的值与α无关。

定理7.2.8 如果A具有相容次序且其对角元全不为零，又矩阵$B=I-\mathrm{diag}(a_{ii}^{-1})A=D^{-1}L+D^{-1}U$，则

1) 若μ是B的p重特征值，则$-\mu$也是B的p重特征值；

2) 若λ是J_ω的特征值，则存在B的一个特征值μ，使

$$\lambda+\omega-1=\omega\mu\lambda^{\frac{1}{2}},\qquad(7.2.19)$$

3) 如果μ是B的特征值，λ满足(7.2.19)，则λ是J_ω的一个特征值。

证明 因为A具有相容次序，所以$B-I$也具有相容次序。根据定理7.2.7的推论，取$\alpha=\pm 1$，$k=\mu$，则

$$\det(B-\mu I)=\det(-B-\mu I)=(-1)^n\det(B+\mu I),$$

从而B的特征多项式可写为

$$\det(B-\mu I)=(-1)^n\mu^{n-2l}(\mu^2-\mu_1^2)(\mu^2-\mu_2^2)\cdots(\mu^2-\mu_l^2),$$

其中μ_i是B的非零特征值，于是结论1)成立。根据(7.2.12)

$$J_\omega-\lambda I=(I-\omega D^{-1}L)^{-1}[\lambda\omega D^{-1}L+\omega D^{-1}U+(1-\omega-\lambda)I],$$

所以

$$\det(J_\omega-\lambda I)=\det[\lambda\omega D^{-1}L+\omega D^{-1}U+(1-\omega-\lambda)I].\quad(7.2.20)$$

因为A具有相容次序且对角元全不为零，根据定理7.2.7的推论可得

$$\begin{aligned}\det(J_\omega-\lambda I)&=(\omega\lambda^{\frac{1}{2}})^n\det\Big(\lambda^{\frac{1}{2}}D^{-1}L+\lambda^{-\frac{1}{2}}D^{-1}U\\&\qquad+\frac{1-\omega-\lambda}{\omega\lambda^{1/2}}I\Big)\\&=(\omega\lambda^{\frac{1}{2}})^n\det\Big(D^{-1}L+D^{-1}U\\&\qquad-\frac{\lambda+\omega-1}{\omega\lambda^{1/2}}I\Big).\end{aligned}\quad(7.2.21)$$

设λ是J_ω的特征值，若$\lambda\neq 0$，则$\frac{\lambda+\omega-1}{\omega\lambda^{1/2}}$是$B$的特征值，记为$\mu$，于是

$$\lambda+\omega-1=\omega\mu\lambda^{\frac{1}{2}}.$$

如果$\lambda=0$，由(7.2.20)得出$\det(J_\omega)=\det(\omega D^{-1}U+(1-\omega)I)=0$，因而必有$1-\omega=0$，即$\omega=1$。故对任何$\lambda$值，(7.2.19)均成立，因此结论2)成立。

设μ是$B=D^{-1}L+D^{-1}U$的特征值，且λ满足(7.2.19)。若$\lambda\neq 0$，则$\mu=(\lambda+\omega-1)/\omega\lambda^{\frac{1}{2}}$，于是由(7.2.21)得到

$$\det(J_\omega-\lambda I)=0$$

即λ是J_ω的特征值；如果$\lambda=0$，由(7.2.19)得到$\omega=1$，于是由(7.2.20)可得$\det(J_\omega)=0$，即$\lambda=0$是J_ω的特征值，从而结论3)成立。

定理7.2.9　如果A是对称正定矩阵，具有相容次序，并且$B=I-D^{-1}A=D^{-1}L+D^{-1}U$，则

1)　B的特征值均为实数且$S(B)<1$

$$2)\ S(J_\omega)=\begin{cases}\left\{\dfrac{\omega S(B)+[\omega^2 S(B)^2-4(\omega-1)]^{\frac{1}{2}}}{2}\right\}^2, & 0<\omega\leqslant\bar{\omega}\\ \omega-1, & \bar{\omega}\leqslant\omega<2\end{cases}\tag{7.2.22}$$

这里$\bar{\omega}$是SOR法的最佳松弛因子，记为

$$\omega_{opt}=\bar{\omega}=\frac{2}{1+\sqrt{1-S(B)^2}},\tag{7.2.23}$$

$$S(J_{\bar{\omega}})=\bar{\omega}-1.\tag{7.2.24}$$

证明　由于A是对称正定矩阵，所以$I-D^{\frac{1}{2}}AD^{-\frac{1}{2}}$也为对称矩阵。又由于

$$B=I-D^{-1}A=D^{-\frac{1}{2}}(I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}})D^{\frac{1}{2}},$$

所以B与$I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$相似，从而$B$的特征值均为实数。

根据定理7.2.8，当μ是B的特征值时，$-\mu$也是B的特征值，而且当λ满足

$$\lambda+\omega-1=\omega|\mu|\lambda^{\frac{1}{2}}$$

时，λ是J_ω的特征值. 设$\lambda_1^{\frac{1}{2}}$和$\lambda_2^{\frac{1}{2}}$为上式的两个根，据定理7.1.1和定理7.2.2，当$0<\omega<2$时，$|\lambda_1|^{\frac{1}{2}}$和$|\lambda_2|^{\frac{1}{2}}$均小于1. 由线性代数知识可得

$$|\mu|<1,\ 即S(B)<1.$$

当μ是B的特征值时，方程式(7.2.19)的两个根

$$\lambda_1^{\frac{1}{2}} = \frac{\omega|\mu| + [\omega^2\mu^2 - 4(\omega-1)]^{\frac{1}{2}}}{2}, \tag{7.2.24}$$

$$\lambda_2^{\frac{1}{2}} = \frac{\omega|\mu| - [\omega^2\mu^2 - 4(\omega-1)]^{\frac{1}{2}}}{2} \tag{7.2.25}$$

均为J_ω的特征值，而且

当$\mu=0$时，$\lambda_1=\lambda_2=1-\omega$；

当$\omega^2\mu^2-4(\omega-1)<0$时，$\lambda_1$与$\lambda_2$为一对共轭复数且$|\lambda_1|=|\lambda_2|=|\omega-1|$；

当$\omega^2\mu^2-4(\omega-1)\geqslant 0$时，$\lambda_1$与$\lambda_2$是正实数且$\lambda_1\lambda_2=(\omega-1)^2$。

因为我们关心的是当ω取什么值时$S(J_\omega)$最小，所以只要对$\omega\in(0, 2)$内的每个值，取出λ_1与λ_2中按模最大的一个来讨论。显然，对于任意实数μ和ω均有

$$|\lambda_1|\geqslant|\lambda_2|,\quad |\lambda_1|\geqslant|\omega-1|,\quad |\lambda_2|\leqslant|\omega-1|. \tag{7.2.26}$$

从(7.2.24)可以看出，对于任意给定的ω，$|\lambda_1|$是$|\mu|$的单调增函数。因此，取$\mu=S(B)$时，$|\lambda_1|$达到最大值$S(J_\omega)$，若记$S(B)=S_b$，则

$$S(J_\omega) = \left|\frac{\omega S_b + [\omega^2 S_b^2 - 4(\omega-1)]^{1/2}}{2}\right|^2. \tag{7.2.27}$$

(7.2.27)中

$$\omega^2 S_b^2 - 4(\omega-1) = 0 \tag{7.2.28}$$

的根为

$$\omega_{1,2} = \frac{2(1\pm\sqrt{1-S_b^2})}{S_b^2}.$$

因为$S_b<1$，所以方程(7.2.28)在区间(0, 2)内仅有一个根，记为$\bar{\omega}$，则

$$\bar{\omega} = \frac{2(1-\sqrt{1-S_b^2})}{S_b^2} = \frac{2}{1+(1-S_b^2)^{\frac{1}{2}}}$$

$$= 1 + \left(\frac{S_b}{1+(1-S_b^2)^{\frac{1}{2}}}\right)^2. \tag{7.2.29}$$

因此，当$0<\omega<\bar{\omega}$时，(7.2.28)的左边取正值，于是

$$S(J_\omega) = \left(\frac{\omega S_b + (\omega^2 S_b^2 - 4(\omega-1))^{\frac{1}{2}}}{2}\right)^2; \tag{7.2.30}$$

当$\bar{\omega}\leqslant\omega<2$时，(7.2.28)的左边取负值，从而(7.2.27)右边$|\cdot|$内取复数，其模为$(\omega-1)^{\frac{1}{2}}$，则

$$S(J_\omega) = \omega - 1, \tag{7.2.31}$$

因此(7.2.22)成立。

(7.2.22)建立了$S(J_\omega)$与ω的函数关系，现在来求$S(J_\omega)$的最小值。首先讨论$0<\omega<\bar{\omega}$的情况，此时$S(J_\omega)$取(7.2.22)的第1式，若令

$$p(\omega) = [\omega^2 S_b^2 - 4(\omega-1)]^{1/2},$$

则

$$p'(\omega) = \frac{\omega S_b^2 - 2}{p(\omega)},$$

$$\frac{d}{d\omega}S(J_\omega) = \frac{1}{2}(\omega S_b + p(\omega))(S_b + p'(\omega))$$

$$= \frac{1}{2}(\omega S_b + p(\omega))\frac{4(S_b^2-1)}{p(\omega)[p(\omega)S_b + 2 - \omega S_b^2]}$$

$$= \begin{cases} <0, & 0<\omega<\bar{\omega}\text{时}; \\ -\infty, & \omega\to\bar{\omega}-0\text{时}. \end{cases} \tag{7.2.32}$$

$$\frac{d^2}{d\omega^2}S(J_\omega) = \frac{1}{2}\left[(S_b + p'(\omega))^2 + (\omega S_b + p(\omega))\frac{S_b^2 - (p'(\omega))^2}{p(\omega)}\right]$$

$$= \frac{S_b + p'(\omega)}{2p(\omega)}(2p(\omega)S_b + \omega S_b^2 - \omega S_b p'(\omega))$$

$$= \begin{cases} <0, & \text{当}0<\omega<\bar{\omega}\text{时}; \\ -\infty, & \text{当}\omega\to\bar{\omega}-0\text{时}. \end{cases}$$

由此可见，当 $0<\omega<\bar{\omega}$ 时，$S(J_\omega)$ 是单调下降函数；当 $\bar{\omega}<\omega<2$ 时，$S(J_\omega)=\omega-1$，是单调上升函数。因此，当 $\omega=\bar{\omega}$ 时，$S(J_\omega)$ 达到最小值，故

$$\omega_{opt}=\bar{\omega}=\frac{2}{1+\sqrt{1-(S(B))^2}}.$$

显然，当 $\omega=\omega_{opt}=\bar{\omega}$ 时，$S(J_\omega)=\bar{\omega}-1=\omega_{opt}-1$。

松弛因子 ω 与 $S(J_\omega)$ 之间的关系，以及 ω_{opt} 的位置见图 7.1。

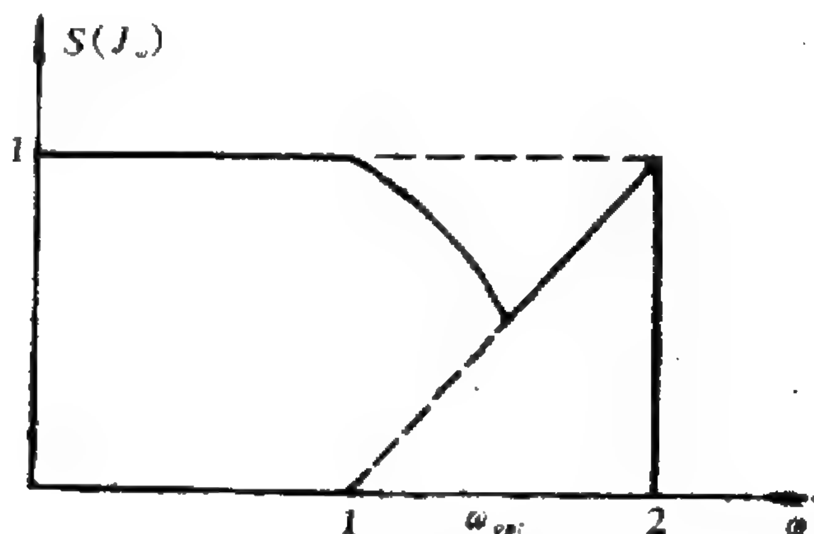


图7.1 超松弛法的谱半径

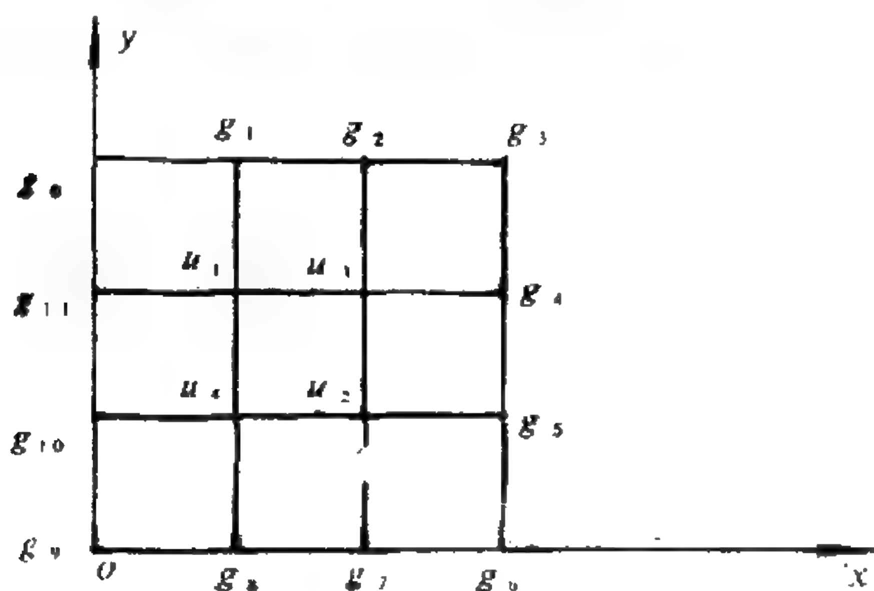
由图 7.1 可以看出，当 ω 从 ω_{opt} 的左边逐渐增加而趋近于 ω_{opt} 时，$S(J_\omega)$ 的切线斜率趋于 $-\infty$，即切线趋于垂直于 ω 轴。但当 ω 从 ω_{opt} 的右边逐渐减小而趋近于 ω_{opt} 时，$S(J_\omega)$ 的切线斜率不变（恒为 1），这就是说 $\omega_{opt}-\Delta\omega$ 与 $\omega_{opt}+\Delta\omega$ 中，后者对应的谱半径将较之前者为小。所以，在实际计算 ω_{opt} 的近似值时，取 ω 略大于 ω_{opt} 比取 ω 略小于 ω_{opt} 更为有利。

例7.2.4.

$$\begin{cases}u_{xx}(x,y)+u_{yy}(x,y)=0,\quad 0<x,y<1,\\ u|_\Gamma=g(x,y)=0.\end{cases}\tag{7.2.32}$$

其中 Γ 是边长为 1 的正方形边界，其真解显然为 $u(x,y)\equiv 0$。

在单位正方形上，用边长$h=1/3$作出均匀的正方形网格，并对网格结点进行编号，如图7.2所示。



7.2 网格

记 $u_1=u\left(\frac{1}{3},\ \frac{2}{3}\right)$，$u_2=u\left(\frac{2}{3},\ \frac{1}{3}\right)$，$u_3=u\left(\frac{2}{3},\ \frac{2}{3}\right)$，$u_4=u\left(\frac{1}{3},\ \frac{1}{3}\right)$，设$u_i$的近似值为$w_i(i=1,2,3,4)$，求$w_i$。精度要求为$\|\mathbf{w}^{(m)}-\mathbf{w}^*\|_\infty\leqslant10^{-7}$。

解　将方程(7.2.32)离散化后，u_i的近似值w_i可以表示为如下形式：

$$
\begin{cases}
w_1=\frac{1}{4}(w_3+w_4+g_1+g_{11}), \\
w_2=\frac{1}{4}(w_3+w_4+g_5+g_7), \\
w_3=\frac{1}{4}(w_1+w_2+g_2+g_4), \\
w_4=\frac{1}{4}(w_1+w_2+g_8+g_{10}).
\end{cases}
\tag{7.2.33}
$$

又因为$g_i=0$，所以(7.2.33)可以写成如下形式

$$\mathbf{A}\mathbf{w}=0.$$

其中

$$A=\begin{pmatrix}1 & 0 & -\frac{1}{4} & -\frac{1}{4}\\ 0 & 1 & -\frac{1}{4} & -\frac{1}{4}\\ -\frac{1}{4} & -\frac{1}{4} & 1 & 0\\ -\frac{1}{4} & -\frac{1}{4} & 0 & 1\end{pmatrix},\quad \boldsymbol{w}=\begin{pmatrix}w_1\\ w_2\\ w_3\\ w_4\end{pmatrix},$$

$$B=I-D^{-1}A=\frac{1}{4}\begin{pmatrix}0 & 0 & 1 & 1\\ 0 & 0 & 1 & 1\\ 1 & 1 & 0 & 0\\ 1 & 1 & 0 & 0\end{pmatrix}.$$

因为

$$\det(\mu I-B)=\mu^2(\mu^2-\frac{1}{4})=0,$$

所以B的特征值为$\mu_1=\mu_2=0$，$\mu_3=\frac{1}{2}$，$\mu_4=-\frac{1}{2}$，则$S(B)=\frac{1}{2}$；又因为A的特征值为$\lambda_1=\lambda_2=1$，$\lambda_3=1-\frac{1}{2}=\frac{1}{2}$，$\lambda_4=1+\frac{1}{2}=1\frac{1}{2}$，所以$A$为对称正定矩阵；据例7.2.3知，$A$具有相容次序，故使用SOR法的最佳松弛因子$\omega_{opt}$可选为

$$\omega_{opt}=\frac{2}{1+\sqrt{1-(S(B))^2}}\doteq 1.072.$$

取迭代初值为：$w_1^{(0)}=0.222222$，$w_2^{(0)}=0.222222$，$w_3^{(0)}=444444$，$w_4^{(0)}=0.111111$。使用SOR法计算，看松弛因子ω对迭代次数的影响。实际迭代次数见下表。

ω	迭代次数	ω	迭代次数
1.020	11	1.072	8
1.040	10	1.202	10
1.052	9	1.402	17

在实际计算中，由于$S(B)$一般也是难于事先求出的，于是ω_{opt}的值无法事先确定。因此，在计算中，如何确定ω_{opt}的近似值是一个比较困难的问题。这里介绍一种在使用SOR法迭代的过程中顺便求出ω_{opt}近似值的方法。

当$\omega \in (0, \omega_{opt})$时，由(7.2.22)的第1式可以求出$S(B)$用$S(J_\omega)$和$\omega$表示的形式

$$S(B)=\frac{S(J_\omega)+(\omega-1)}{\omega(S(J_\omega))^{\frac{1}{2}}}. \qquad (7.2.34)$$

因此，若能求出$S(J_\omega)$，$S(B)$也就知道了，同时就能求出ω_{opt}的近似值。

首先取一个不太大的正数ω，使之满足$0<\omega\leqslant\omega_{opt}$，然后使用SOR迭代格式

$$x^{(m)}=J_\omega x^{(m-1)}+g$$

进行迭代。于是有

$$\begin{aligned} x^{(m)}-x^{(m-1)} &= J_\omega(x^{(m-1)}-x^{(m-2)})=\cdots \\ &= J_\omega^{m-1}(x^{(1)}-x^{(0)}). \end{aligned} \qquad (7.2.35)$$

当J_ω是具有线性初等因子的实矩阵，并且$S(J_\omega)$是J_ω的按模最大的单重特征值时，由于

$$\frac{\|x^{(m)}-x^{(m-1)}\|_\infty}{\|x^{(m-1)}-x^{(m-2)}\|_\infty}=\lambda^{(m)}\to S(J_\omega), \text{ 当} m\to\infty\text{时} \qquad (7.2.36)$$

当迭代经过若干步迭代$\lambda^{(m)}$逐渐稳定下来时，就把$\lambda^{(m)}$作为$S(J_\omega)$的近似值，再代入(7.2.34)求出$S(B)$，然后利用(7.2.23)求出ω_{opt}的近似值。再利用这个ω_{opt}代替原来的ω继续进行迭代。直到找到比较满意的ω_{opt}的近似值为止。使用这个方法，由于难于判断$S(J_\omega)$的近似值是否已较好地接近真值，以及利用乘幂法求单个最大特征值也要受到有关条件和误差的影响，要想得到较好的ω_{opt}的近似值，往往需要很多次的迭代。

对于某些特殊问题，矩阵 B 的按模最大的特征值 μ 可以使用某些办法估计出来，这时就可以按(7.2.23)算出 ω_{opt} 的近似值，直接用到SOR迭代格式中去。但是应该注意，对于具有相容次序的矩阵在求得 ω_{opt} 的近似值后，应该采用比它稍大一点的值，作为实际计算中的松弛因子。

对于那些目前还没有理论公式可以计算最佳松弛因子的矩阵，我们可以用试算的办法来确定它的近似值。最简单的办法是，从同一初始向量出发，用不同的松弛因子 ω 迭代相同的次数，然后比较它们相应的剩余或误差，并选取使剩余或误差之模最小的松弛因子作为最佳松弛因子的近似值。这个方法既简单，又有效。特别，当使用者需要多次求解具有相同系数矩阵的方程组时更是如此。

有时还可以根据求解问题中最佳松弛因子的性质，利用迭代过程中的信息，将它估计出来，而在估算过程中应尽量不要占用过多的计算时间。

§3 预处理和块松弛法

在使用迭代方法之前，特别是当系数阵 A 的条件数很大时，为了加快迭代过程的收敛速度或者为了适应迭代矩阵所具有的性质的需要，有时要对原方程组进行事先处理，即预处理。

在使用 SOR 方法时，求新的近似解是按分量逐个地进行的，对于某些方程组，如能适当进行预处理，确定新的近似解则可以成组地进行，有时还能加快收敛的速度。

3.1 预处理法

设线性方程组为

$$Ax=b. \tag{7.3.1}$$

其中A为对称正定阵，为了讨论方便设A的主对角线元均为1，则A能分解成下列形式

$$A=I-L-L^T.$$

于是方程组

$$\begin{cases} By=d, \\ B=(I-\omega L)^{-1}A(I-\omega L)^{-T}, \\ y=(I-\omega L^T)x,\quad d=(I-\omega L)^{-1}b \end{cases} \tag{7.3.2}$$

与(7.3.1)等价。

使用迭代法求解时，方程组(7.3.2)与(7.3.1)的迭代矩阵是不同的，其谱半径一般也不相同，因此它们的收敛速度也不一样。如果对变形后的方程组(7.3.2)使用简单迭代法，其迭代格式为

$$y^{(k+1)}=d+(I-B)y^{(k)}, \tag{7.3.3}$$

它与原方程组使用简单迭代法计算时，其步骤及迭代效果都不同。由于B是三个矩阵的乘积，由(7.3.3)从$y^{(k)}$计算$y^{(k+1)}$时可经下列的代换把计算过程分解为更基本的运算。首先，令

$$By^{(k)}=z^{(k)}, \tag{7.3.4}$$

则

$$y^{(k+1)}=d+y^{(k)}-z^{(k)}, \tag{7.3.5}$$

于是

$$z^{(k)}=(I-\omega L)^{-1}A(I-\omega L)^{-T}y^{(k)}. \tag{7.3.6}$$

由于

$$(I-\omega L)^{-T}y^{(k)}=x^{(k)}, \tag{7.3.7}$$

故

$$(I-\omega L)z^{(k)}=Ax^{(k)}=F^{(k)}. \tag{7.3.8}$$

(7.3.4)～(7.3.8)的过程可以用下列算法来描述。

算法7.3.1. (预处理算法)

1) 输入 ω、$\boldsymbol{x}^{(0)}$，置 $\boldsymbol{y}^{(0)}=(I-\omega L)^T\boldsymbol{x}^{(0)}$；

2) 由 $(I-\omega L)\boldsymbol{d}=\boldsymbol{b}$ 计算 $\boldsymbol{d}$；

3) 对 $k=0,1,\cdots,k_{max}$ 做

 3.1) 置 $\boldsymbol{F}^{(k)}=A\boldsymbol{x}^{(k)}$，

 3.2) 由 $(I-\omega L)\boldsymbol{z}^{(k)}=\boldsymbol{F}^{(k)}$ 计算 $\boldsymbol{z}^{(k)}$，

 3.3) $\boldsymbol{y}^{(k+1)}=\boldsymbol{d}+\boldsymbol{y}^{(k)}-\boldsymbol{z}^{(k)}$；

 3.4) 由 $(I-\omega L)^T\boldsymbol{x}^{(k+1)}=\boldsymbol{y}^{(k+1)}$ 计算 $\boldsymbol{x}^{(k+1)}$，

 3.5) 如 $\boldsymbol{x}^{(k+1)}$ 可接受则转4)，

 3.6) NEXTk。

4) 结束。

这里 k_{max} 是允许的最大的迭代次数，ω 是预选的某一适当的因子。

定理7.3.1 若矩阵 A 是主对角线元素为1的对称正定矩阵，则当 $\omega=1$ 时，迭代法(7.3.3)收敛。

证明 要证(7.3.3)收敛，只需证 $S(I-B)<1$。设矩阵 B 的特征值为 μ，其相应的特征向量为 p，则

$$Bp=\mu p,$$

即

$$(I-\omega L)^{-1}A(I-\omega L)^{-T}p=\mu p。$$

两边用 p^T 左乘，得到

$$p^T(I-\omega L)^{-1}A(I-\omega L)^{-T}p=\mu p^Tp, \qquad (7.3.9)$$

令

$$v=(I-\omega L)^{-T}p,$$

由(7.3.9)得到

$$v^TAv=\mu p^Tp,$$

$$\mu=\frac{v^TAv}{p^Tp}=\frac{v^TAv}{v^T(I-\omega L)(I-\omega L^T)v}$$

$$=\frac{v^TAv}{(1-\omega)v^Tv+\omega v^TAv+\omega^2v^TLL^Tv}, \qquad (7.3.10)$$

故当 $\omega=1$ 时，

$$\mu=\frac{v^TAv}{v^TAv+v^TLL^Tv}. \tag{7.3.11}$$

由于 A 为对称正定矩阵，且 $v^TLL^Tv=v^TL(v^TL)^T\geqslant 0$，所以 $0<\mu\leqslant 1$。又因矩阵 $I-B$ 的特征值为 $1-\mu$，所以

$$S(I-B)<1.$$

实际上，ω 的最优值常是大于 1 的，但这里参数 ω 的取值不象 SOR 法那样严格，通常可用 $\omega=1$ 作为近似最优值。

对于一般的实对称正定矩阵 A，可以将它分解成

$$A=D-\bar{L}-\bar{L}^T,$$

其中 $D=\mathrm{diag}(a_{11},a_{22},\cdots,a_{nn})$。于是方程组

$$\begin{cases} By=d, \\ B=(D^{\frac{1}{2}}-\omega L)^{-1}A(D^{\frac{1}{2}}-\omega L)^{-T}, \\ y=(D^{\frac{1}{2}}-\omega L^T)x, \\ d=(D^{\frac{1}{2}}-\omega L)^{-1}b, \end{cases} \tag{7.3.12}$$

其中 $L=\bar{L}D^{-\frac{1}{2}}$ 与方程组 $Ax=b$ 等价。同样可对(7.3.12)使用简单迭代法，其迭代格式与(7.3.3)相同

$$y^{(k+1)}=d+(I-B)y^{(k)}.$$

仿算法 7.3.1，不难写出一般的含有实对称正定矩阵 A 的方程组 $Ax=b$ 的预处理算法。

算法7.3.1 （一般的实对称正定方程组的预处理算法）

1）输入 ω，$x^{(0)}$，置 $y^{(0)}=(D^{\frac{1}{2}}-\omega L^T)x^{(0)}$；

2）由 $(D^{\frac{1}{2}}-\omega L)d=b$，计算 d；

3）对 $k=0,1,\cdots,k_{max}$ 做

3.1）$w^{(k)}=Ax^{(k)}$，

3.2） 由$(D^{\frac{1}{2}}-\omega L)\boldsymbol{z}^{(k)}=\boldsymbol{w}^{(k)}$计算$\boldsymbol{z}^{(k)}$，

3.3） $\boldsymbol{y}^{(k+1)}=\boldsymbol{d}+\boldsymbol{y}^{(k)}-\boldsymbol{z}^{(k)}$，

3.4） 由$(D^{\frac{1}{2}}-\omega L^T)\boldsymbol{x}^{(k+1)}=\boldsymbol{y}^{(k+1)}$求$\boldsymbol{x}^{(k+1)}$，

3.5） 如$\boldsymbol{x}^{(k+1)}$可接受则转4)，

3.6） NEXTk，

4） 结束。

当A的条件数cond(A)较大时。可以通过预处理法来降低条件数。当$\mathbf{A}$是具有“性质$\mathbf{A}$”的对称正定矩阵时，可以证明，一定存在ω(如$\omega=0$)，使cond$(B)\leqslant$cond$(\mathbf{A})$。

如果将$\mathbf{A}$进行三角分解

$$\mathbf{A}=\bar{L}\bar{L}^T,$$

令

$$D^{\frac{1}{2}}-\omega L=\bar{L},$$

则(7.3.12)中的矩阵和向量分别换成

$$B=\bar{L}^{-1}\mathbf{A}\bar{L}^{-T},\qquad \boldsymbol{y}=\bar{L}^T\boldsymbol{x},\qquad \boldsymbol{d}=\bar{L}^{-1}\boldsymbol{b}.$$

于是B的条件数cond$(B)=1$，此时经一次迭代就得到精确解。但是求$\bar{L}^{-1}$相当于求$\mathbf{A}^{-1}$，因此，寻找容易求逆的近似三角分解矩阵$\bar{L}$，是一个引人注意的问题。J. A. Meijerink 和 H.A; Van der Vorst 于 1977 年提出的不完全分解法是很有价值的。将该法用于某些问题的处理，现已取得了很好的效果。

3.2 块松弛法

分块迭代法在确定新的近似分量时是成组进行的。先将求解方程组中的方程和未知数进行分组，使每个方程和每个未知数属于且仅属于一个组。进行迭代时，从每个方程组中同时获得的一组近似解的分量，如此按组进行计算，直到求得全部新的近似解分量为止。

在建立块迭代格式时，分组的方法是可以根据需要来选定

的。下面介绍一种自然数序号分组法。首先将自然数集$W=\{1, 2,\cdots,n\}$分成互不相交的子集：$R_1, R_2, \cdots, R_m$，$R_1\cup R_2\cup\cdots\cup R_m=W$且$R_i\cap R_j=0(i\neq j)$。对于两个分组法，当且仅当分成的子集的个数和相应子集均相等时，才称这两个分组法相同。如果把分组法用Π_i表示，对于$W=\{1,2,3\}$的分组就有：

Π_0：$R_1=\{1\}$，　　$R_2=\{2\}$，　　$R_3=\{3\}$；

Π_1：$R_1=\{1,2\}$，　　$R_2=\{3\}$；

Π_2：$R_1=\{1,3\}$，　　$R_2=\{2\}$；

Π_3：$R_1=\{2,1\}$，　　$R_2=\{3\}$；

Π_4：$R_1=\{3\}$，　　$R_2=\{1,2\}$。

根据分组法的规定，此时$\Pi_1=\Pi_3$，但$\Pi_1\neq\Pi_4$。通常我们把分组法：$R_k=\{k\}$，$k=1,2,\cdots,n$，叫做Π_n。

如果方程组

$$Ax=b$$

经过Π分组后，变成

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_m \end{pmatrix} = \begin{pmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \vdots \\ \bar{b}_m \end{pmatrix}, \qquad (7.3.12)$$

其中子块$A_{rs}(r,s=1,2,\cdots,m)$是矩阵A中行号属于子集R_r中的行，以及列号属于R_s中的列相交叉点的元素按原顺序排列而成的子块。子列向量$\bar{x}_s$和$\bar{b}_s$分别是列向量x和b中分量的序号属于R_s中的分量按原顺序排列而成的列子向量。经过这样分块后，方程组(7.3.12)可以写成

$$\sum_{s=1}^{m} A_{rs}\bar{x}_s=\bar{b}_r \qquad (r=1,2,\cdots,m)。 \qquad (7.3.13)$$

例如，方程组

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}. \tag{7.3.14}$$

按分组法 Π 分为：$R_1=\{1,3\}$，$R_2=\{2\}$ 时，方程组(7.3.14)可写成(7.3.13)的形式为

$$\begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} + \begin{pmatrix} a_{12} \\ a_{32} \end{pmatrix} [x_2] = \begin{pmatrix} b_1 \\ b_3 \end{pmatrix}, \tag{7.3.15}$$

$$[a_{21}\ a_{23}] \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} + [a_{22}][x_2] = [b_2].$$

如果方程组(7.3.13)中矩阵 $A_{rr}(r=1,\cdots,m)$ 均为非奇异，则可以仿照简单迭代法的形式写出块简单迭代格式

$$A_{rr}\bar{x}_r^{(k+1)} = \bar{b}_r - \sum_{\substack{s=1 \\ s\neq r}}^{m} A_{rs}\bar{x}_s^{(k)}, \quad (r=1,2,\cdots,m) \tag{7.3.16}$$

或

$$\bar{x}_r^{(k+1)} = \sum_{s=1}^{m} B_{rs}\bar{x}_s^{(k)} + \bar{c}_r\ (r=1,2,\cdots,m), \tag{7.3.17}$$

其中

$$B_{rs} = \begin{cases} -A_{rr}^{-1}A_{rs}, & \text{当} r \neq s, \\ 0, & \text{当} r = s, \end{cases} \tag{7.3.18}$$

$$\bar{c}_r = A_{rr}^{-1}\bar{b}_r. \tag{7.3.19}$$

我们也可以将(7.3.17)写成矩阵形式

$$x^{(k+1)} = B^{(\Pi)}x^{(k)} + F^{(\Pi)}. \tag{7.3.20}$$

其中

$$B^{(\Pi)} = (D^{(\Pi)})^{-1}E^{(\Pi)}, \quad \text{而} F^{(\Pi)} = (D^{(\Pi)})^{-1}b,$$

$$E^{(\Pi)} = D^{(\Pi)} - A, \quad D^{(\Pi)} = \mathrm{diag}(A_{11},\ A_{22},\ \cdots,\ A_{mm}).$$

完全仿照逐次松弛法(7.1.5)的形式，同样可以写出块松弛法的计算公式

$$A_{rr}\bar{x}_r^{(k+1)} = \omega(\bar{b}_r - \sum_{s=1}^{r-1} A_{rs}\bar{x}_s^{(k+1)} - \sum_{s=r+1}^{m} A_{rs}\bar{x}_s^{(k)})$$

$$-(1-\omega)A_{rr}\bar{\mathbf{x}}_r^{(k)} \tag{7.3.21}$$

$$(r=1,2,\cdots,m)$$

或
$$\bar{\mathbf{x}}_r^{(k+1)}=\omega(\mathbf{c}_r-\sum_{s=1}^{r-1}B_{rs}\bar{\mathbf{x}}_s^{(k+1)}-\sum_{s=r+1}^{m}B_{rs}\bar{\mathbf{x}}_s^{(k)})$$

$$-(1-\omega)I_r\bar{\mathbf{x}}_r^{(k)}. \tag{7.3.22}$$

$$(r=1,2,\cdots,m)$$

将(7.3.22)写成矩阵形式，得

$$\mathbf{x}^{(k+1)}=J_\omega^{(II)}\mathbf{x}^{(k)}+(I-\omega L^{(II)})^{-1}\omega\mathbf{F}^{(II)}, \tag{7.3.23}$$

其中

$$J_\omega^{(II)}=(I-\omega L^{(II)})^{-1}(\omega U^{(II)}+(1-\omega)I), \tag{7.3.24}$$

$$L^{(II)}=(D^{(II)})^{-1}A_L^{(II)}, \tag{7.3.25}$$

$$U^{(II)}=(D^{(II)})^{-1}A_U^{(II)}, \tag{7.3.26}$$

而

$$A_L^{(II)}=\begin{pmatrix} 0 & & & \\ A_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ A_{m1} & \cdots & A_{m\,m-1} & 0 \end{pmatrix},$$

$$A_U^{(II)}=\begin{pmatrix} 0 & A_{12} & \cdots & A_{1m} \\ & \ddots & \ddots & \vdots \\ & & \ddots & A_{m-1\,m} \\ & & & 0 \end{pmatrix}.$$

例如，设

$$\mathbf{A}=\begin{pmatrix} 4 & -1 & 0 & -1 & 0 \\ -1 & 4 & -1 & 0 & -1 \\ 0 & -1 & 4 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 \\ 0 & -1 & 0 & -1 & 4 \end{pmatrix},$$

Π_1: $R_1=\{1,2,3\}$, $R_2=\{4,5\}$,

则

$$A_{11}=\begin{pmatrix}4 & -1 & 0\\ -1 & 4 & -1\\ 0 & -1 & 4\end{pmatrix},\qquad A_{12}=\begin{pmatrix}-1 & 0\\ 0 & -1\\ 0 & 0\end{pmatrix},$$

$$A_{21}=\begin{bmatrix}-1 & 0 & 0\\ 0 & -1 & 0\end{bmatrix},\qquad A_{22}=\begin{bmatrix}4 & -1\\ -1 & 4\end{bmatrix},$$

$$D^{(\Pi_1)}=\begin{bmatrix}A_{11} & 0\\ 0 & A_{22}\end{bmatrix},\qquad A_L^{(\Pi_1)}=\begin{bmatrix}0 & 0\\ A_{21} & 0\end{bmatrix},$$

$$A_U^{(\Pi_1)}=\begin{bmatrix}0 & A_{12}\\ 0 & 0\end{bmatrix},$$

$$(D^{(\Pi_1)})^{-1}=\begin{pmatrix}\frac{15}{56} & \frac{4}{56} & \frac{1}{56} & 0 & 0\\ \frac{4}{56} & \frac{16}{56} & \frac{4}{56} & 0 & 0\\ \frac{1}{56} & \frac{4}{56} & \frac{15}{56} & 0 & 0\\ 0 & 0 & 0 & \frac{4}{15} & \frac{1}{15}\\ 0 & 0 & 0 & \frac{1}{15} & \frac{4}{15}\end{pmatrix},$$

$$L^{(\Pi_1)}=\begin{pmatrix}0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ \frac{4}{15} & \frac{1}{15} & 0 & 0 & 0\\ \frac{1}{15} & \frac{4}{15} & 0 & 0 & 0\end{pmatrix},$$

$$U^{(\Pi_1)} = \begin{pmatrix} 0 & 0 & 0 & \frac{16}{56} & \frac{4}{56} \\ 0 & 0 & 0 & \frac{4}{56} & \frac{16}{56} \\ 0 & 0 & 0 & \frac{1}{56} & \frac{4}{56} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

利用块松弛法求解，在计算$\bar{x}_1^{(k+1)}$时需要解一个低阶方程组(7.3.21)。通常可以用消去法求解。但是，由于A_{rr}不随k变化，一个有效的办法是，先把A_{ii}进行三角分解，並保存起来，反复利用它来求解(7.3.21)中的$\bar{x}_r^{(k+1)}$　$(k=1,2,\cdots,m)$如此所需要完成的计算量就可以减少。

完全仿照定理7.2.2的证明方法，可以证明如下定理。

定理7.3.2　若A为实对称正定矩阵，则当$0<\omega<2$时，块松弛法(7.3.21)收敛。

对于n阶方阵A，使用Π分组法将$W=\{1, 2, \cdots, n\}$分成m组后，我们定义m阶矩阵$\chi=(\chi_{rs})_{m\times m}$，其中

$$\chi_{rs} = \begin{cases} 0, & \text{当 } A_{rs}=0; \\ 1, & \text{当 } A_{rs}\neq 0. \end{cases} \qquad (7.3.27)$$

定义7.3.1　若矩阵χ具有“性质A”，则称矩阵A具有性质$A^{(\Pi)}$。

定义7.3.2　若矩阵χ具有相容次序，则称矩阵A具有Π相容次序。

如果$\Pi=\Pi_0$时χ有相容次序，那末矩阵A就具有通常所称的相容次序。

当矩阵A是具有Π相容次序的对称正定阵且$B^{(\Pi)}=I-$

$(D^{(\Pi)})^{-1}A$，则可以仿照定理7.2.7的推论、定理7.2.8和定理7.2.9的方法证明如下结果：

1）行列式 $\Delta=\det(\alpha A_L^{(\Pi)}+\alpha^{-1}A_U^{(\Pi)}-kD^{(\Pi)})$ 的值与 α 无关，其中 $\alpha\neq 0$，k 为任意值。

2）将 B 换成 $B^{(\Pi)}$，J_ω 换成 $J_\omega^{(\Pi)}$ 后，有类似于定理7.2.8和 $S(B^{(\Pi)})<1$ 的结果，从而可以得到块松弛法的最佳松弛因子为

$$\omega_{opt}^{(\Pi)}=\frac{2}{1+(1-(S(B^{(\Pi)}))^2)^{1/2}},\tag{7.3.28}$$

而且

$$S(J_\omega(\overset{(\Pi)}{\underset{opt}{E}}))=\omega_{opt}^{(\Pi)}-1.\tag{7.3.29}$$

在实际问题中，有些问题导出的方程组的系数矩阵A往往具有块三对角的形式

$$A=\begin{pmatrix} A_{11} & A_{12} & & & \\ A_{21} & A_{22} & A_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & & & A_{m-1\,m} \\ & & & A_{m\,m-1} & A_{mm} \end{pmatrix},$$

这种形式的矩阵显然具有 Π 相容次序。因此，只要块三对角阵具有对称正定性，那末，逐次松弛法的全部结论都适用于块松弛法。

块松弛法也是一种解线性方程组的有效方法.特别当A_{rr}不太复杂，而且先对 A_{rr} 进行分解的工作量不太大时，用该法往往能提高收敛速度。而且，对于那些没有"性质A"但是经适当分块后可以具有性质 $A^{(\Pi)}$的矩阵，使用块松弛法进行求解更为合适。然而尽管松弛法有时可获得较好的效果，可是当A_{rr}较为复杂或者求解程序效率不高时，往往并不能使总的计算时间

减少。以上这些问题，在使用块松弛法时也必须考虑，以便根据问题的具体情况决定是否选用块松弛法。

§4 Chebyshev 半迭代法

对于一阶线性定常迭代法$x^{(m)} = Bx^{(m-1)} + g$，如果收敛速度很慢，则实际意义并不大。我们希望找到一种方法能加快它的收敛速度。基于一阶线性定常迭代法的 Chebyshev 加速，当迭代矩阵B的特征值均为实数时，往往能取得较好的效果。

4.1 一般加速原则

给定一个收敛很慢，甚至不收敛的实数序列：x_0，x_1，…，x_n，…，记为$\{x_i\}$，我们可以由它来构造一个新序列：y_0，y_1，…，y_n，…，记为$\{y_i\}$，使新序列$\{y_i\}$的收敛比$\{x_i\}$的收敛更快。例如

$$
\begin{aligned}
y_0 &= x_0, \\
y_1 &= \frac{1}{2}(x_0 + x_1), \\
y_2 &= \frac{1}{3}(x_0 + x_1 + x_2), \\
&\vdots
\end{aligned}
\tag{7.4.1}
$$

可以证明，若$\{x_i\}$收敛，那末$\{y_i\}$一定收敛；甚至当$\{x_i\}$不收敛时，$\{y_i\}$也可能收敛。如果$\{y_i\}$收敛，我们称$\{x_i\}$为 Cesàro 可求和。

构造新序列的更一般的原则，可以先考虑下列三角形列阵：

$$
\begin{array}{llll}
\alpha_{00} & & & \\
\alpha_{10} & \alpha_{11} & & \\
\alpha_{20} & \alpha_{21} & \alpha_{22} & \\
\cdots & \cdots & \cdots & \cdots
\end{array}
$$

其中

$$\sum_{k=0}^{m} \alpha_{mk} = 1, \qquad m = 0, 1, 2, \cdots. \tag{7.4.2}$$

再与$\{x_i\}$生成新序列$\{y_i\}$，其中

$$y_i = \sum_{k=0}^{i} \alpha_{ik} x_k \qquad i = 0, 1, 2, \cdots \tag{7.4.3}$$

显然，当对所有的 i 均有 $\alpha_{ik} = 0$（$k<i$ 时）而 $\alpha_{ii} = 1$ 时，$\{y_i\}$ 就是 $\{x_i\}$；当 $\alpha_{ik} = 1/(i+1)$ $(k = 0, 1, \cdots, i)$ 时，(7.4.3)就是(7.4.1)。

求解线性方程组 $\boldsymbol{Ax} = \boldsymbol{b}$ 的一阶线性定常迭代法(7.1.9)所确定的向量序列 $\boldsymbol{x}^{(0)}$，$\boldsymbol{x}^{(1)}$，$\cdots$收敛于方程组 $\boldsymbol{Ax} = \boldsymbol{b}$ 的解 $\boldsymbol{x}^*$。今以以下方式定义新向量序列$\{\boldsymbol{u}^{(m)}\}$：

$$\boldsymbol{u}^{(m)} = \sum_{k=0}^{m} \alpha_{mk} \boldsymbol{x}^{(k)}. \tag{7.4.4}$$

其中系数 α_{mk} 满足条件（7.4.2）。由此所定义的过程称为关于一阶线性定常迭代法(7.1.9)的一种半迭代法。

在条件(7.4.2)的限制下，我们总可以找到适当的系数，使$\{\boldsymbol{u}^{(m)}\}$收敛于方程组 $\boldsymbol{Ax} = \boldsymbol{b}$ 的解 $\boldsymbol{x}^*$。为了分析半迭代法的收敛性，令

$$\begin{cases} \boldsymbol{\varepsilon}^{(m)} = \boldsymbol{x}^{(m)} - \boldsymbol{x}^*, \\ \boldsymbol{\eta}^{(m)} = \boldsymbol{u}^{(m)} - \boldsymbol{x}^*. \end{cases} \tag{7.4.5}$$

从(7.4.4)和(7.4.2)可以推得

$$\begin{aligned} \boldsymbol{\eta}^{(m)} = \boldsymbol{u}^{(m)} - \boldsymbol{x}^* &= \sum_{k=0}^{m} \alpha_{mk} \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \sum_{k=0}^{m} \alpha_{mk} \\ &= \sum_{k=0}^{m} \alpha_{mk} (\boldsymbol{x}^{(k)} - \boldsymbol{x}^*) = \sum_{k=0}^{m} \alpha_{mk} \boldsymbol{\varepsilon}^{(k)}. \end{aligned}$$

由(7.1.12)得到

$$\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{B}^k \boldsymbol{\varepsilon}^{(0)} = \boldsymbol{B}^k \boldsymbol{\eta}^{(0)},$$

因此

$$\eta^{(m)} = \sum_{k=0}^{m} \alpha_{mk} B^k \varepsilon^{(0)} = p_m(B)\eta^{(0)}, \tag{7.4.6}$$

其中

$$p_m(B) = \sum_{k=0}^{m} \alpha_{mk} B^k \text{ 且 } p_m(1) = 1. \tag{7.4.7}$$

为了使 $\{u^{(m)}\}$ 收敛于 x^*，即 $\eta^{(m)} \to 0$，就需要 $\lim\limits_{m\to\infty} p_m(B) = 0$。满足这些要求的多项式 $p_m(x)$ 是存在的，例如 $p_m(x) = x^m$，此时 $p_m(B) = B^m$，由于 $S(B) < 1$，自然就有 $\lim\limits_{m\to\infty} p_m(B) = 0$，且 $p_m(1) = 1^m = 1$，即 $u^{(m)} \to x^*$。那末，应该选取怎样的多项式(7.4.7)才能使向量序列 $\{u^{(m)}\}$ 的收敛最快？为此，必须首先弄清收敛快慢的度量。

类似于§1，我们首先假定 B 具有线性无关的特征向量系 $x_1, x_2, \cdots, x_n$，而 $\lambda_1, \lambda_2, \cdots, \lambda_n$ 是相应的特征值，于是 $\eta^{(0)}$（即 $\varepsilon^{(0)}$）可以表示为

$$\begin{aligned}
\eta^{(0)} &= a_1 x + a_2 x + \cdots + a_n x_n, \\
\eta^{(m)} &= p_m(B)\eta^{(0)} = a_1 p_m(B) x_1 + \cdots + a_n p_m(B) x_n \\
&= p_m(\lambda_1) a_1 x_1 + p_m(\lambda_2) a_2 x_2 + \cdots + p_m(\lambda_n) a_n x_n.
\end{aligned} \tag{7.4.8}$$

由此可以看出，$\eta^{(m)}$ 收敛于零的速度是由 $\max\limits_{1\le i\le n} |p_m(\lambda_i)|$ 决定的。为使 $\{u^{(m)}\}$ 收敛最快，只需选取所有 m 阶多项式 $p_m(x)$ 中使下式达到极小值者

$$\begin{cases} S(p_m(B)) = \max\limits_{1\le i\le n} |p_m(\lambda_i)|, \\ p_m(1) = 1, \end{cases}$$

因而，可以类似于定义7.1.2一样定义渐近收敛速度。

定义7.4.1 称 $R_\infty(p_m(B)) = \lim\limits_{m\to\infty}\left(-\frac{1}{m}\ln(S(p_m(B)))\right)$ 为半迭代法(7.4.4)的渐近收敛速度。

当 $p_m(x)=x^m$ 时，$R_\infty(p_m(B))=R(B)$，因此，定义 7.4.1 是定义7.1.2的自然推广。

4.2 Chebyshev 半迭代法

假定B的全部特征值λ_i都是实数，并且满足如下条件

$$\alpha\leqslant\lambda_i\leqslant\beta<1 \quad (\alpha<\beta), \tag{7.4.9}$$

其中α可以小于 -1。在这种条件下，用 Chebyshev 多项式构成半迭代法能取得加快收敛速度的效果。通常将这样的半迭代法称为 Chebyshev 半迭代法。由于λ_i事先不知道，所以 $\max\limits_{1\leqslant i\leqslant n}|p_m(\lambda_i)|$的极小值也难于求出。通常用求

$$\begin{cases}\max\limits_{\alpha\leqslant\lambda\leqslant\beta}|p_m(\lambda)|, \\ p_m(1)=1\end{cases} \tag{7.4.10}$$

的极小化问题来代替。当然有

$$\max_{1\leqslant i\leqslant n}|p_m(\lambda_i)|^{\frac{1}{m}}\leqslant\max_{\alpha\leqslant\lambda\leqslant\beta}|p_m(\lambda)|^{\frac{1}{m}}。 \tag{7.4.11}$$

今引入新的变量

$$\gamma=\gamma(\lambda)=\frac{2\lambda-(\alpha+\beta)}{\beta-\alpha}。 \tag{7.4.12}$$

显然有$\gamma(\alpha)=-1$，$\gamma(\beta)=1$，而且当λ满足(7.4.9)时，$-1\leqslant\gamma\leqslant1$。令

$$z=\gamma(1)=\frac{2-(\alpha+\beta)}{\beta-\alpha}。 \tag{7.4.13}$$

又因为

$$2-(\alpha+\beta)>2-(2\beta)=2(1-\beta)>0。$$

$$\beta-\alpha>0。$$

$$2-(\alpha+\beta)-(\beta-\alpha)=2-2\beta>0。$$

所以

$$z>1。$$

由(7.4.12)解出$\lambda=\frac{1}{2}[(\beta-\alpha)\gamma+\beta+\alpha]$代入$p_m(\lambda)$得到

$$p_m(\lambda)=p_m\left(\frac{(\beta-\alpha)\gamma+\beta+\alpha}{2}\right)=Q_m(\gamma). \qquad (7.4.14)$$

因此

$$\max_{\alpha\leqslant\lambda\leqslant\beta}|p_m(\lambda)|=\max_{-1\leqslant\gamma\leqslant1}|Q_m(\gamma)|$$

于是极小化问题(7.4.10)变成

$$\begin{cases}\max\limits_{-1\leqslant\gamma\leqslant1}|Q_m(\gamma)|=\min\\ Q_m(z)=1\end{cases}. \qquad (7.4.15)$$

这个问题的解答是

$$Q_m(\gamma)=T_m(\gamma)/T_m(z). \qquad (7.4.16)$$

其中$T_m(\gamma)$就是m阶 Chebyshev 多项式，其参数表达式可以写成

$$\begin{cases}x=\cos\theta,\\ T_m(x)=\cos m\theta.\end{cases} \qquad 0\leqslant\theta\leqslant\pi.$$

这里

$$T_m(x)=\cos(m\ \arccos x)$$
$$=\frac{(x+\sqrt{x^2-1})^m+(x-\sqrt{x^2-1})^m}{2}, \qquad (7.4.17)$$

它对$x>1$时也是有意义的。由(7.4.15)的解可得(7.4.10)的解为

$$p_m(\lambda)=Q_m(\gamma)=Q_m\left(\frac{2\lambda-(\beta+\alpha)}{\beta-\alpha}\right)$$
$$=T_m\left(\frac{2\lambda-(\beta+\alpha)}{\beta-\alpha}\right)\Big/T_m\left(\frac{2-(\beta+\alpha)}{\beta-\alpha}\right), \qquad (7.4.18)$$

而且

$$\min\left(\max_{\alpha\leqslant\lambda\leqslant\beta}|p_m(\lambda)|\right)=\max_{\alpha\leqslant\lambda\leqslant\beta}\left|T_m\left(\frac{2\lambda-(\beta+\alpha)}{\beta-\alpha}\right)\Big/T_m(z)\right|.$$
$$=\frac{1}{T_m(z)} \qquad (7.4.19)$$

有了 $p_m(\lambda)$ 的表达式(7.4.18)后，利用 Chebyshev 多项式 $T_m(x)$ 的递推性质

$$\begin{cases} T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x) & (m=1,2,\cdots), \\ T_0(x) = 1, \ T_1(x) = x, \end{cases} \tag{7.4.20}$$

再通过计算 $p_m(\lambda) = \alpha_{m0} + \alpha_{m1}\lambda' + \cdots + \alpha_{mm}\lambda^m$ 的系数 α_{mi} 就可以构造 $\boldsymbol{u}^{(m)}$。因为

$$T_0(x) = 1, \qquad T_1(x) = x, \qquad T_2(x) = 2x^2 - 1,$$

可以求得

$$p_0(\lambda) = \alpha_{00} = 1, \qquad \boldsymbol{u}^{(0)} = \boldsymbol{x}^{(0)};$$

$$p_1(\lambda) = \alpha_{10} + \alpha_{11}\lambda = T_1\left(\frac{2\lambda - (\beta+\alpha)}{\beta - x}\right) \Big/ T_1(z)$$

$$= \frac{2\lambda - (\beta+\alpha)}{\beta - \alpha} \cdot \frac{\beta - \alpha}{2 - (\beta+\alpha)} = \frac{2\lambda - (\beta+\alpha)}{2 - (\beta+\alpha)}.$$

经比较系数，得

$$\alpha_{10} = -\frac{\beta+\alpha}{2-(\beta+\alpha)}, \qquad \alpha_{11} = \frac{2}{2-(\beta+\alpha)}.$$

于是

$$\boldsymbol{u}^{(1)} = \frac{-(\beta+\alpha)}{2-(\beta+\alpha)}\boldsymbol{x}^{(0)} + \frac{2}{2-(\beta+\alpha)}\boldsymbol{x}^{(1)}$$

$$= \frac{2}{2-(\beta+\alpha)}(B\boldsymbol{x}^{(0)} + \boldsymbol{g}) - \frac{(\beta+\alpha)}{2-(\beta+\alpha)}\boldsymbol{x}^{(0)}. \tag{7.4.21}$$

如此继续进行，就可以求出 $\boldsymbol{u}^{(m)}$ 来。但是，这种算法比较麻烦，下面我们导出利用 $\boldsymbol{u}^{(m-1)}$，$\boldsymbol{u}^{(m)}$ 来计算 $\boldsymbol{u}^{(m+1)}$ 的递推关系式。从(7.4.18)可得

$$p_m(B) = T_m\left(\frac{2B - (\beta+\alpha)I}{\beta - \alpha}\right) \Big/ T_m(z),$$

由(7.4.6)和(7.4.20)可以推出

$$\boldsymbol{\eta}^{(m+1)}=\left[T_{m+1}\left(\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right)\Big/T_{m+1}(z)\right]\boldsymbol{\varepsilon}^{(0)}$$

$$=\left\{\left[2\left(\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right)T_m\left(\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right)\right.\right.$$

$$\left.\left.-T_{m-1}\left(\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right)\right]\cdot\left[T_{m+1}(z)\right]^{-1}\right\}\boldsymbol{\varepsilon}^{(0)}$$

$$=2\left[\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right]\frac{T_m(z)}{T_{m+1}(z)}\boldsymbol{\eta}^{(m)}$$

$$-\frac{T_{m-1}(z)}{T_{m+1}(z)}\boldsymbol{\eta}^{(m-1)}, \qquad (7.4.22)$$

将 $\boldsymbol{\eta}^{(m)}=\boldsymbol{u}^{(m)}-\boldsymbol{x}^*$ 代入上式后得到

$$\boldsymbol{u}^{(m+1)}=2\left[\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right]\frac{T_m(z)}{T_{m-1}(z)}\boldsymbol{u}^{(m)}$$

$$-\frac{T_{m-1}(z)}{T_{m+1}(z)}\boldsymbol{u}^{(m-1)}+\left[I-2\left(\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right)\right.$$

$$\left.\cdot\frac{T_m(z)}{T_{m+1}(z)}+\frac{T_{m-1}(z)}{T_{m+1}(z)}I\right]\boldsymbol{x}^*. \qquad (7.4.23)$$

但是

$$\left[I-2\left(\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right)\frac{T_m(z)}{T_{m+1}(z)}+\frac{T_{m-1}(z)}{T_{m+1}(z)}I\right]\boldsymbol{x}^*$$

$$=2\left[zI-\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right]\frac{T_m(z)}{T_{m+1}(z)}\boldsymbol{x}^*$$

$$=\frac{4(I-B)}{\beta-\alpha}\frac{T_m(z)}{T_{m+1}(z)}\boldsymbol{x}^*=\frac{4}{\beta-\alpha}\frac{T_m(z)}{T_{m+1}(z)}\boldsymbol{g}, \qquad (7.4.24)$$

因此，当 $m\geqslant1$时，

$$u^{(m+1)} = 2\left[\frac{2B-(\beta+\alpha)I}{\beta-\alpha}\right]\frac{T_m(z)}{T_{m+1}(z)}u^{(m)} - \frac{T_{m-1}(z)}{T_{m+1}(z)}u^{(m-1)}$$

$$+\frac{4}{\beta-\alpha}\frac{T_m(z)}{T_{m+1}(z)}g, \tag{7.4.25}$$

由(7.4.21)和(7.4.25)可以求出$\{u^{(m)}\}$。如果令$\rho_1=1$,

$$\rho_m = 2zT_{m-1}(z)/T_m(z), \quad m=2,3,\cdots \tag{7.4.26}$$

(7.4.25)还可以简化成

$$u^{(m+1)} = \frac{\rho_{m+1}}{2-(\beta+\alpha)}\left\{[2B-(\beta+\alpha)I]u^{(m)}+2g\right\}$$

$$+(1-\rho_{m+1})u^{(m-1)}.$$

将$T_{m+1}(x)=2xT_m(x)-T_{m-1}(x)$的两边乘上$1/2xT_m(x)$得到

$$\frac{T_{m+1}(x)}{2xT_m(x)} = 1-\frac{T_{m-1}(x)}{2xT_m(x)} = 1-\frac{1}{4x^2}\frac{2xT_{m-1}(x)}{T_m(x)}.$$

由(7.4.26)可得

$$\rho_{m+1}^{-1} = 1-\frac{1}{4z^2}\rho_m,$$

而

$$\rho_2 = \frac{2zT_1(z)}{T_2(z)} = \frac{2z^2}{2z^2-1},$$

从而得到基于一阶线性定常迭代法(7.1.9)的 Chebyshev 迭代法的简化计算公式为

$$\begin{cases} \text{初始值} \quad x^{(0)}=u^{(0)}, \\ u^{(1)} = \dfrac{2}{2-(\alpha+\beta)}(Bx^{(0)}+g) - \dfrac{\alpha+\beta}{2-(\alpha+\beta)}x^{(0)}, \\ u^{(m+1)} = \dfrac{\rho_{m+1}}{2-(\alpha+\beta)}\{[2B-(\alpha+\beta)I]u^{(m)}+2g\} \\ \qquad +(1-\rho_{m+1})u^{(m-1)}. \end{cases} \tag{7.4.27}$$

$$\rho_1 = 1, \qquad \rho_2 = \frac{2z^2}{2z^2-1},$$

其中，
$$\rho_{m+1} = \left[1 - \frac{1}{4z^2}\rho_m\right]^{-1}, \qquad m = 2, 3, \cdots$$

$$z = \frac{(\alpha+\beta)}{\beta-\alpha}.$$

从这些计算公式中可以看出，Chebyshev 半迭代法是一种二阶线性非定常迭代法。使用该迭代法时要注意参数 α 和 β 的选取，在一般情况下可运用特征值估计的有关方法来定 α 和 β。

下面我们来讨论 Chebyshev 半迭代法的收敛性。

由(7.4.17)可得

$$T_m(z) = \frac{(z+\sqrt{z^2-1})^m + (z-\sqrt{z^2-1})^m}{2}$$

$$= \frac{(1+\sqrt{1-1/z^2})^m + (1-\sqrt{1-1/z^2})^m}{(2/z^m)}.$$

令 $\sigma = 1/z$，则

$$T_m(z) = \frac{(1+\sqrt{1-\sigma^2})^m + (1-\sqrt{1-\sigma^2})^m}{2\sigma^m}$$

$$= \frac{1}{2}\left[\left(\frac{1+\sqrt{1-\sigma^2}}{\sigma}\right)^m + \left(\frac{1-\sqrt{1-\sigma^2}}{\sigma}\right)^m\right].$$

令 $\tau = \sigma/(1+\sqrt{1-\sigma^2})$，则

$$T_m(z) = \frac{1}{2}\left[\frac{1}{\tau^m} + \tau^m\right] = \frac{1+\tau^{2m}}{2\tau^m},$$

从而推得

当 $z>1$ 时，$[T_m(z)]^{-1} = \dfrac{2\tau^m}{1+\tau^{2m}}$，

其中

$$\tau = \frac{\sigma}{1+\sqrt{1-\sigma^2}}, \qquad \sigma = \frac{1}{z}.$$

显然，当 $m\to\infty$ 时，$[T_m(z)]^{-1}\to 0$。于是根据(7.4.8)和(7.4.19)可以推出

当 $m\to\infty$ 时， $\eta^{(m)}=u^{(m)}-x^*\to 0$，

即 $u^{(m)}\to x^*$。因此，不管一阶线性定常迭代法(7.1.9)是否收敛，对于满足上述条件的 Chebyshev 半迭代法是一定收敛的。

如果原迭代法收敛，它的渐近收敛速度为 $R(B)$，Chebyshev 半迭代法的渐近收敛速度为 $R(p_m(B))$，那末，两者收敛速度的快慢如何呢？下面将作出比较。

据定义7.4.1，$R(p_m(B))=\lim\limits_{m\to\infty}\left(-\dfrac{1}{m}\ln S(p_m(B))\right)$，由(7.4.11)可以推出

$$S(p_m(B))\leqslant(T_m(z))^{-1}$$

从而

$$\begin{aligned}R(p_m(B))&\geqslant\lim_{m\to\infty}\left(-\frac{1}{m}\ln\frac{1}{T_m(z)}\right)\\&=\lim_{m\to\infty}\left(-\frac{1}{m}\ln\frac{2\tau^m}{1+\tau^{2m}}\right)=-\ln\tau.\end{aligned}\quad(7.4.28)$$

另一方面，我们可以作出线性函数 $\bar{p}_1(x)=\alpha_1x+\alpha_0$，其中 $\alpha_0+\alpha_1=1$，使

$$\max_{\alpha\leqslant\lambda\leqslant\beta}|\bar{p}_1(\lambda)|\leqslant\max_{1\leqslant i\leqslant n}|\lambda_i|.$$

这里 λ_i 和 $\bar{p}_1(\lambda_i)$ 分别是 B 和 $\bar{p}_1(B)$ 的特征值。由此可得

$$R(B)\leqslant R(\bar{p}_1(B))\leqslant-\ln\frac{1}{T_1(z)}=-\ln\sigma.$$

因此，

$$\frac{R(p_m(x))}{[R(B)]^{1/2}}\geqslant\frac{-\ln\tau}{[-\ln\sigma]^{1/2}},$$

其中 $\tau=\dfrac{\sigma}{1+\sqrt{1-\sigma^2}}$，$0<\sigma<1$。

现在我们来证明

$$\lim_{\sigma\to 1}\frac{-\ln\tau}{[-\ln\sigma]^{1/2}}=\sqrt{2}.$$

令 $\sigma = e^{-a}$，则 $a = -\ln\sigma$，由于

$$-\ln\frac{\sigma}{1+\sqrt{1-\sigma^2}} = -\frac{1}{2}\ln\frac{1-\sqrt{1-\sigma^2}}{1+\sqrt{1-\sigma^2}}$$

$$\geqslant (1-\sigma^2)^{\frac{1}{2}} = (1-e^{-2a})^{\frac{1}{2}}$$

$$= e^{-a}(e^{2a}-1)^{\frac{1}{2}} \geqslant \sigma(2a)^{\frac{1}{2}}$$

$$= \sigma(-2\ln\sigma)^{\frac{1}{2}},$$

并且

$$-\ln\frac{\sigma}{1+\sqrt{1-\sigma^2}} = a+\ln(1+\sqrt{1-\sigma^2}) \leqslant a+(1-\sigma^2)^{\frac{1}{2}}$$

$$= a+(1-e^{-2a})^{\frac{1}{2}} \leqslant a+(1-(1-2a))^{\frac{1}{2}}$$

$$= -\ln\sigma+(-2\ln\sigma)^{\frac{1}{2}},$$

所以，

$$\frac{\sigma(-2\ln\sigma)^{1/2}}{(-2\ln\sigma)^{1/2}} \leqslant \frac{-\ln\dfrac{\sigma}{1+\sqrt{1-\sigma^2}}}{(-2\ln\sigma)^{1/2}} \leqslant \frac{-\ln\sigma+(-2\ln\sigma)^{\frac{1}{2}}}{(-2\ln\sigma)^{1/2}}.$$

由于 $\sigma\to1^-$ 时，

$$\frac{\sigma(-2\ln\sigma)^{\frac{1}{2}}}{(-2\ln\sigma)^{\frac{1}{2}}}\to1, \qquad \frac{-\ln\sigma+(-2\ln\sigma)^{1/2}}{(-2\ln\sigma)^{1/2}}\to1,$$

故

$$\lim_{\sigma\to1^-}\frac{-\ln\dfrac{\sigma}{1+\sqrt{1-\sigma^2}}}{(-2\ln\sigma)^{1/2}} = 1,$$

即

$$\lim_{\sigma\to1^-}\frac{-\ln\tau}{(-\ln\sigma)^{1/2}} = \sqrt{2}.$$

这就是说，当 σ 接近于 1 时，Chebyshev 半迭法的渐近收敛速度比原来的一阶线性定常迭代法 (7.1.9) 的渐近收敛速度

快得多。因此，当 B 的特征值全部是实数时，采用 Chebyshev 半迭代法来加速迭代过程是有效的。

4.3 简单迭代法的 Chebyshev 加速

前面我们讨论了一般的一阶线性定常迭代法的 Chebyshev 加速问题。如果满足条件的某种迭代法已经给出，问题就转化为如何选择其中的参数了。我们以简单迭代法为例来说明这个问题。

设线性方程组为 $A\boldsymbol{x}=\boldsymbol{b}$，其系数矩阵 A 是对称正定的，其相应的简单迭代法为：

$$\boldsymbol{x}^{(m-1)}=G\boldsymbol{x}^{(m)}+\boldsymbol{c}, \qquad (7.4.29)$$

其中 $G=I-D^{-1}A$，　$A=\mathrm{diag}(a_{ii})$，　$\boldsymbol{c}=D^{-1}\boldsymbol{b}$；

因为

$$\bar{G}=D^{\frac{1}{2}}GD^{-\frac{1}{2}}=I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$$

且 $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 是对称正定阵，所以 G 的特征值 μ 均为实数且小于 1，于是存在实数 α，β 使

$$\alpha\leqslant\mu\leqslant\beta<1.$$

据(7.4.27)，简单迭代法的 Chebyshev 半迭法可以表为

$$\boldsymbol{u}^{(m+1)}=\frac{\rho_{m+1}}{2-(\alpha+\beta)}\left\{[2G-(\beta+\alpha)I]\boldsymbol{u}^{(m)}+2\boldsymbol{c}\right\}+(1-\rho_{n+1})\boldsymbol{u}^{(m-1)}. \qquad (7.4.30)$$

在(7.4.30)中，G 的特征值进一步满足某些特殊条件时，计算格式变得更为简单些。

1) 当 G 的所有特征值均为小于 1 的相等实数，即 $\alpha=\beta$ 时，显然，可取 $\rho_m=1$，$(m=1,2,\cdots)$ 从而(7.4.30)变成

$$\boldsymbol{u}^{(m+1)}=\frac{1}{1-\alpha}[(G-\alpha I)\boldsymbol{u}^{(m)}+\boldsymbol{c}]. \qquad (7.4.31)$$

又因为 $S(G-\alpha I)=0$，所以迭代格式(7.4.31)的收敛速度是非常快的。

2）如果矩阵 A 是具有相容次序的对称正定阵，此时 $S(G)<1$，故可设 $-\alpha=\beta=S(G)$，于是

$$\tau=\frac{2-(\alpha+\beta)}{\beta-\alpha}=\frac{2}{2S(G)}=\frac{1}{S(G)}. \tag{7.4.32}$$

将(7.4.32)代入(7.4.27)得到

$$\rho_1=1, \qquad \rho_2=2/(2-(S(G))^2),$$

$$\rho_{m+1}=\left(1-\frac{1}{4}(S(G))^2\rho_m\right)^{-1} \qquad (m=2,3,\cdots), \tag{7.4.33}$$

而(7.4.30)变成

$$u^{(m+1)}=\rho_{m+1}(Gu^{(m)}+c)+(1-\rho_{m+1})u^{(m-1)}. \tag{7.4.34}$$

因此，在这种条件下，Chebyshev 半迭代法的计算公式可以由(7.4.34)和(7.4.33)表示。

根据上面的讨论和(7.4.28)，半迭代法的渐近收敛速度应满足

$$R_\infty(p_m(G))\geqslant -\ln\tau, \quad \tau=\frac{S(G)}{1+\sqrt{1-(S(G))^2}},$$

即

$$R_\infty(p_m(G))\geqslant -\ln\frac{S(G)}{1+\sqrt{1-(S(G))^2}}$$

$$=-\frac{1}{2}\ln\frac{1-\sqrt{1-(S(G))^2}}{1+\sqrt{1-(S(G))^2}}. \tag{7.4.35}$$

如果改用逐次松弛法，并取最佳松弛因子为 ω_{opt}，其渐近收敛速度为

$$R_\infty(J_{\omega_{opt}})=-\ln(\omega_{opt}-1)=-\ln\left(\frac{2}{1+\sqrt{1-(S(G))^2}}-1\right)$$

$$=-\ln\frac{1-\sqrt{1-(S(G))^2}}{1+\sqrt{1-(S(G))^2}}. \tag{7.4.36}$$

当矩阵A是具有相容次序的对称正定矩阵时，比较(7.4.35)和(7.4.36)，可以看出，SOR法的渐近收敛速度为基于简单迭代法的Chebyshev半迭代法的两倍，因此，当线性方程组的系数矩阵A是具有相容次序的对称正定阵时，一般说来，采取逐次松弛法是较为有利的。

§5 非线性迭代法

非线性迭代法的特点是新近似解是已有近似解的非线性函数。本节只介绍解线性方程组的最速下降法和共轭斜量法等两种非线性迭代法。它适用于系数矩阵为对称正定阵的情况。由于该法不需要选取参数，所以使用比较方便。共轭斜量法，实质上是一种直接法，如果没有舍入误差，对于n阶方程组最多迭代n步便可求得精确解。但是共轭斜量法又具有迭代的计算公式，所以就其形式来说又是一种迭代法。它还可以直接应用于解非线性方程组和求函数的极值问题，因此，近年来受到广泛的重视。它与其它迭代法相比较，不足的是，存贮量较大和计算过程中对舍入误差比较敏感。

5.1 线性方程组的等价问题

设有线性方程组

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \qquad (i=1,2,\cdots,n) \tag{7.5.1}$$

或$Ax=b$，其中系数矩阵A是对称正定的。

考虑下列二次函数的极小值问题

$$F(x) = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_ix_j - \sum_{i=1}^{n} b_ix_i$$

$$= \frac{1}{2}(Ax,x) - (b,x). \tag{7.5.2}$$

这里A仍为对称正定矩阵，$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^T$。如果我们把$n$维向量看成是$n$维空间$R^n$中的一点，那末，我们可以证明下列定理。

定理 7.5.1 向量$\boldsymbol{x}^*$为线性方程组（7.5.1）的解的充要条件是，$\boldsymbol{x}^*$是极小值问题（7.5.2）的极小点。即

$$F(\boldsymbol{x}^*)=\min_{\boldsymbol{x}\in R^n}F(\boldsymbol{x}).$$

证明 因为$a_{ij}=a_{ji}$，所以

$$\frac{\partial F}{\partial x_i}=\sum_{j=1}^{n}a_{ij}x_j-b_i=-r_i,$$

即
$$\operatorname{grad}F(\boldsymbol{x})=A\boldsymbol{x}-\boldsymbol{b}=-\boldsymbol{r}. \tag{7.5.3}$$

也就是说，将任意向量$\boldsymbol{x}$代入方程(7.5.1)中所得到的剩余向量为$\boldsymbol{r}=\boldsymbol{b}-A\boldsymbol{x}$，则$-\boldsymbol{r}$就是函数$F(\boldsymbol{x})$在$\boldsymbol{x}$处的梯度向量。若$\boldsymbol{x}^*$使函数$F(\boldsymbol{x})$取极小值，即$F(\boldsymbol{x}^*)=\min\limits_{\boldsymbol{x}\in R^n}F(\boldsymbol{x})$，则

$$\operatorname{grad}F(\boldsymbol{x})\big|_{\boldsymbol{x}^*}=A\boldsymbol{x}^*-\boldsymbol{b}=0,$$

即$\boldsymbol{x}^*$是线性方程组(7.5.1)的解。

反之，若$\boldsymbol{x}^*$是线性方程组（7.5.1）的解，而A又是对称正定阵，故有

$$\begin{aligned}F(\boldsymbol{x})-F(\boldsymbol{x}^*)&=\frac{1}{2}(A\boldsymbol{x},\boldsymbol{x})-(\boldsymbol{b},\boldsymbol{x})-\frac{1}{2}(A\boldsymbol{x}^*,\boldsymbol{x}^*)\\&\quad+(\boldsymbol{b},\boldsymbol{x}^*)\\&=\frac{1}{2}[(A\boldsymbol{x},\boldsymbol{x})-2(\boldsymbol{b},\boldsymbol{x})+(A\boldsymbol{x}^*,\boldsymbol{x}^*)]\\&\quad-(A\boldsymbol{x}^*-\boldsymbol{b},\boldsymbol{x}^*)\\&=\frac{1}{2}[(A\boldsymbol{x},\boldsymbol{x})-2(A\boldsymbol{x}^*,\boldsymbol{x})+(A\boldsymbol{x}^*,\boldsymbol{x}^*)]\\&=\frac{1}{2}(A(\boldsymbol{x}-\boldsymbol{x}^*),\boldsymbol{x}-\boldsymbol{x}^*)\geqslant 0,\end{aligned} \tag{7.5.4}$$

即 x^* 是 $F(x)$ 的唯一极小点。

定理表明，求线性方程组(7.5.1)的解的问题，可以转化成求函数 $F(x)$ 的极小点。求 $F(x)$ 的极小点的计算过程，大体可以归结为：从任意向量 x 出发，沿着某一合适的方向向量 p 进行修改，从而得到一个新的近似解向量 $\bar{x}$，使得 $F(x)>F(\bar{x})$，如此不断地修改下去，最终得到 $F(x)$ 的极小点 x^*。从计算的过程可以看出，求 $F(x)$ 的极小点 x^* 的问题，关键是选择修正方向向量。

5.2 最速下降法

最速下降法是一种逐步逼近的方法，每次的逼近值都是从 $F(x)$ 的值的减小最快的方向上选取，由此而命名这个方法。其具体迭代公式推导如下：

首先任取一个向量 $x^{(0)}$ 作为初始向量，选取一个方向 p_0 使 $F(x)$ 的值沿方向 p_0 减小最快，从(7.5.3)知道

$$\operatorname{grad}F(x^{(0)})=Ax^{(0)}-b=-r^{(0)},$$

即 $F(x)$ 从 $x^{(0)}$ 出发沿 $r^{(0)}$ 方向减小最快。所以选取 $p_0=r^{(0)}$，再在 $x^{(0)}+\alpha p^{(0)}=x^{(0)}+\alpha r^{(0)}$ 上选取一点 $x^{(1)}$ 使

$$F(x^{(1)})=\min_{\alpha}F(x^{(0)}+\alpha r^{(0)}).$$

设

$$\begin{aligned}\phi(\alpha)&=F(x^{(0)}+\alpha r^{(0)})\\&=\frac{1}{2}(A(x^{(0)}+\alpha r^{(0)}),\ (x^{(0)}+\alpha r^{(0)}))\\&\quad-(b,x^{(0)}+\alpha r^{(0)})\\&=\frac{1}{2}\alpha^2(Ar^{(0)},r^{(0)})-\alpha(r^{(0)},r^{(0)})\\&\quad+F(x^{(0)}).\end{aligned}$$

从方程

$$\frac{d\phi(\alpha)}{d\alpha}=\alpha(Ar^{(0)},r^{(0)})-(r^{(0)},r^{(0)})=0$$

中解得

$$\alpha=(r^{(0)},r^{(0)})/(Ar^{(0)},r^{(0)})=\alpha_0,$$

又因为 $\frac{d^2\phi(\alpha)}{d\alpha^2}=(Ar^{(0)},r^{(0)})>0$，所以

$$\min_{\alpha}F(x^{(0)}+\alpha r^{(0)})=F(x^{(0)}+\alpha_0 r^{(0)}).$$

于是解的第一次近似为

$$\begin{cases}x^{(1)}=x^{(0)}+\alpha_0 r^{(0)},\\ \alpha_0=(r^{(0)},r^{(0)})/(Ar^{(0)},r^{(0)}),\\ r^{(0)}=b-Ax^{(0)}.\end{cases}\qquad(7.5.5)$$

重复上述过程，如果第 k 次近似向量 $x^{(k)}$ 已经求得，那末，从 $x^{(k)}$ 出发，则使 $F(x)$ 之值减少最快的方向为 $r^{(k)}=b-Ax^{(k)}$，同样可以求出

$$\min_{\alpha}F(x^{(k)}+\alpha r^{(k)})=F(x^{(k)}+\alpha_k r^{(k)}),$$

其中

$$\alpha_k=(r^{(k)},r^{(k)})/(Ar^{(k)},r^{(k)}).$$

于是第$k+1$次近似为

$$\begin{cases}x^{(k+1)}=x^{(k)}+\alpha_k r^{(k)},\\ \alpha_k=\dfrac{(r^{(k)},r^{(k)})}{(Ar^{(k)},r^{(k)})},\quad r^{(k)}=b-Ax^{(k)},\end{cases}\qquad(7.5.6)$$

而且满足如下关系式：

$$\begin{aligned}F(x^{(k+1)})-F(x^{(k)})&=F(x^{(k)}+\alpha_k r^{(k)})-F(x^{(k)})\\&=\frac{1}{2}\alpha_k^2(Ar^{(k)},r^{(k)})-\alpha_k(r^{(k)},r^{(k)})\\&=\frac{1}{2}\frac{(r^{(k)},r^{(k)})^2}{(Ar^{(k)},r^{(k)})}-\frac{(r^{(k)},r^{(k)})^2}{(Ar^{(k)},r^{(k)})}\end{aligned}$$

$$= -\frac{1}{2}\frac{(r^{(k)}, r^{(k)})^2}{(Ar^{(k)}, r^{(k)})} < 0, \tag{7.5.7}$$

即

$$F(x^{(k+1)}) < F(x^{(k)}).$$

这个结果说明了数列$\{F(x^{(k)})\}$是单调下降的。但是，是否有

$$F(x^{(k)}) \to F(x^*),$$

还需要加以证明。为此先证明下列引理。

引理 7.5.1 设正数λ_n，λ_1分别为实对称正定阵 A 的最大与最小特征值，则不等式：

$$F(x^{(k+1)}) - F(x^*) \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^{2(k+1)} [F(x^{(0)}) - F(x^*)] \tag{7.5.8}$$

成立。

证明 根据(7.5.4)和 $Ax^* = b$ 可以证明

$$F(x^{(k)}) - F(x^*) = \frac{1}{2}(A(x^{(k)} - x^*), x^{(k)} - x^*)$$

$$= \frac{1}{2}(Ax^{(k)} - b, A^{-1}(Ax^{(k)} - b))$$

$$= \frac{1}{2}(r^{(k)}, A^{-1}r^{(k)}); \tag{7.5.9}$$

又由(7.5.7)，可得

$$\frac{F(x^{(k)}) - F(x^*)}{F(x^{(k)}) - F(x^{(k+1)})} = \frac{(r^{(k)}, A^{-1}r^{(k)})(Ar^{(k)}, r^{(k)})}{(r^{(k)}, r^{(k)})^2}. \tag{7.5.10}$$

根据线性代数的知识，当 A 为对称正定阵时，A 的特征值可以排成：$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$，而且一定存在一组数$\beta_1, \beta_2, \cdots, \beta_n$ 使(7.5.10)化为如下形式：

$$\frac{(r^{(k)},A^{-1}r^{(k)})(Ar^{(k)},r^{(k)})}{(r^{(k)},r^{(k)})^2}=\frac{\sum\beta_i^2\lambda_i^{-1}\sum\beta_i^2\lambda_i}{(\sum\beta_i^2)^2}$$

$$=\frac{1}{(\sum\beta_i^2)^2}\sum\beta_i^2\frac{\sqrt{\lambda_n\lambda_1}}{\lambda_i}\cdot\sum\beta_i^2\frac{\lambda_i}{\sqrt{\lambda_n\lambda_1}}$$

$$\leqslant\frac{1}{4(\sum\beta_i^2)^2}\left[\sum\beta_i^2\left(\frac{\sqrt{\lambda_n\lambda_1}}{\lambda_i}+\frac{\lambda_i}{\sqrt{\lambda_n\lambda_1}}\right)\right]^2. \qquad (7.5.11)$$

设$\frac{\sqrt{\lambda_n\lambda_1}}{\lambda_i}=y$，则$\sqrt{\frac{\lambda_1}{\lambda_n}}\leqslant y\leqslant\sqrt{\frac{\lambda_n}{\lambda_1}}$。因为$y+\frac{1}{y}$在$y\in\left[\sqrt{\frac{\lambda_1}{\lambda_n}},\sqrt{\frac{\lambda_n}{\lambda_1}}\right]$的最大值只能在$y$的区间端点$\sqrt{\frac{\lambda_1}{\lambda_n}}$或$\sqrt{\frac{\lambda_n}{\lambda_1}}$上取得，而在两端点上取值均为$\sqrt{\frac{\lambda_1}{\lambda_n}}+\sqrt{\frac{\lambda_n}{\lambda_1}}$，从而(7.5.11)可以化成：

$$\frac{(r^{(k)},A^{-1}r^{(k)})(Ar^{(k)},r^{(k)})}{(r^{(k)},r^{(k)})^2}\leqslant\frac{1}{4(\sum\beta_i^2)^2}\left[\sum\beta_i^2\left(\sqrt{\frac{\lambda_1}{\lambda_n}}+\sqrt{\frac{\lambda_n}{\lambda_1}}\right)\right]^2=\frac{(\lambda_1+\lambda_n)^2}{4\lambda_1\lambda_n}=\frac{1}{e}. \qquad (7.5.12)$$

由(7.5.10)和(7.5.12)得到

$$F(x^{(k)})-F(x^{(k+1)})\geqslant e(F(x^{(k)})-F(x^*)).$$

当$\lambda_1\neq\lambda_n$时，$0<e<1$，于是

$$\begin{aligned}F(x^{(k+1)})-F(x^*)&=[F(x^{(k)})-F(x^*)]-[F(x^{(k)})-F(x^{(k+1)})]\\&\leqslant[F(x^{(k)})-F(x^*)]-e[F(x^{(k)}(-F)x^*)]\\&=(1-e)[F(x^{(k)})-F(x^*)].\end{aligned} \qquad (7.5.13)$$

重复使用不等式（7.5.13），最后可得

$$F(x^{(k+1)})-F(x^*)\leqslant(1-e)^{k+1}[F(x^{(0)})-F(x^*)]$$

$$= \left(1 - \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\right)^{k+1} [F(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^*)]$$

$$= \left(\frac{\lambda_n - \lambda_1}{\lambda_1 + \lambda_n}\right)^{2(k+1)} [F(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^*)],$$

于是(7.5.8)得证。

定理 7.5.2 设A为对称正定阵，则

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^{k+1} \sqrt{\frac{2l}{\lambda_1}}. \tag{7.5.14}$$

其中$l = F(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^*)$。

证明 因为 $(A\boldsymbol{x}, \boldsymbol{x}) = \sum \beta_i^2 \lambda_i \geqslant \sum \lambda_1 \beta_i^2 = \lambda_1(\boldsymbol{x}, \boldsymbol{x})$，所以

$$(\boldsymbol{x}, \boldsymbol{x}) \leqslant \frac{1}{\lambda_1}(A\boldsymbol{x}, \boldsymbol{x}),$$

于是

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|^2 = (\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*, \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*)$$

$$\leqslant \frac{1}{\lambda_1}(A(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*), \ \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*)$$

$$= \frac{2}{\lambda_1}(F(\boldsymbol{x}^{(k+1)}) - F(\boldsymbol{x}^*)).$$

再根据引理7.5.1的结果，得到

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\|^2 \leqslant \frac{2}{\lambda_1}\left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^{2(k+1)} (F(\boldsymbol{x}^{(0)}) - F(\boldsymbol{x}^*)),$$

即

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*\| \leqslant \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^{k+1} \sqrt{\frac{2l}{\lambda_1}}.$$

因为$\lambda_n \geqslant \lambda_1 > 0$，所以$0 \leqslant \dfrac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} < 1$，于是当$k \to \infty$时，$x^{(k+1)} \to x^*$，即最速下降法是收敛的。但是，如果$\lambda_n$与$\lambda_1$相差很大，即$(\lambda_n - \lambda_1)/(\lambda_n + \lambda_1) \doteq 1$时，由(7.5.14)知道，最速下降法收敛得很慢。实际计算也表明最速下降法一般收敛较慢，特别对于病态方程组更是如此。因此，必须设法提高最速下降法的收敛速度。

5.3 共轭斜量法

解线性方程组的最速下降法，由于它具有局部性质，即在x附近函数$F(x)$沿方向r下降较快，但是从总体来说并不如此，因此收敛速度并不理想，目前使用得较少。下面我们将选择一些方向，使得$F(x)$沿着这些方向能更快地逼近它的最小值。由于选择的方向具有"共轭"的性质，所以称为**共轭斜量法**。

在构造共轭斜量法的计算公式之前，先引入关于A—共轭（正交）的概念。

设A为对称正定矩阵，x、y是实的列向量，如果

$$(x, Ay) = (Ax, y) = 0, \tag{7.5.15}$$

则称向量x和y是A—共轭（正交）的。

下面推导共轭斜量法的计算公式。

对于任意的初始向量$x^{(0)}$，有剩余向量

$$r^{(0)} = b - Ax^{(0)}.$$

从$x^{(0)}$出发沿$p^{(0)} = r^{(0)}$寻找新的近似向量$x^{(1)}$，取

$$x^{(1)} = x^{(0)} + \alpha_0 r^{(0)},$$

$$\alpha_0 = (r^{(0)}, p^{(0)}) / (p^{(0)}, Ap^{(0)}),$$

再从$x^{(1)}$出发沿$p^{(1)} = r^{(1)} + \beta_0 p^{(0)}$（其中$r^{(1)} = b - Ax^{(1)}$）寻找新的近似向量$x^{(2)}$，并要求$p^{(1)}$与$p^{(0)}$成为$A$—共轭，从而

$$\beta_0 = -(\boldsymbol{r}^{(1)},\ A\boldsymbol{p}^{(0)})/(\boldsymbol{p}^{(0)}, A\boldsymbol{p}^{(0)}).$$

于是 $\boldsymbol{x}^{(2)}$可取为

$$\begin{cases} \boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)} + \alpha_1 \boldsymbol{p}^{(1)}, \\ \alpha_1 = (\boldsymbol{r}^{(1)}, \boldsymbol{p}^{(1)})/(\boldsymbol{p}^{(1)},\ A\boldsymbol{p}^{(1)}). \end{cases}$$

如果 $\boldsymbol{x}^{(i)}$ 已经求得，再去寻找新的近似解向量时，可以从 $\boldsymbol{x}^{(i)}$ 出发选取方向向量 $\boldsymbol{p}^{(i)} = \boldsymbol{r}^{(i)} + \beta_{i-1}\boldsymbol{p}^{(i-1)}$，并要求 $\boldsymbol{p}^{(i)}$与$\boldsymbol{p}^{(i-1)}$成为A—共轭，即

$$(A\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i-1)}) = (\boldsymbol{p}^{(i)}, A\boldsymbol{p}^{(i-1)}) = 0,$$

从而

$$\beta_{i-1} = -(\boldsymbol{r}^{(i)}, A\boldsymbol{p}^{(i-1)})/(\boldsymbol{p}^{(i-1)},\ A\boldsymbol{p}^{(i-1)}), \qquad (7.5.16)$$

其中 $\boldsymbol{r}^{(i)} = \boldsymbol{b} - A\boldsymbol{x}^{(i)}$，于是新的近似解向量 $\boldsymbol{x}^{(i+1)}$可取为

$$\begin{cases} \boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \alpha_i \boldsymbol{p}^{(i)}, \\ \alpha_i = (\boldsymbol{r}^{(i)}, \boldsymbol{p}^{(i)})/(\boldsymbol{p}^{(i)}, A\boldsymbol{p}^{(i)}). \end{cases} \qquad (7.5.17)$$

以上就是共轭斜量法的主要计算步骤，实际计算中，我们还可以将上面的公式进行简化。为此，首先说明向量系$\{\boldsymbol{r}^{(i)}\}$与$\{\boldsymbol{p}^{(i)}\}$之间的关系和构造方法。因为

$$\begin{cases} \boldsymbol{r}^{(i+1)} = \boldsymbol{b} - A\boldsymbol{x}^{(i+1)} = \boldsymbol{b} - A\boldsymbol{x}^{(i)} - \alpha_i A\boldsymbol{p}^{(i)} = \boldsymbol{r}^{(i)} - \alpha_i A\boldsymbol{p}^{(i)}, \\ \boldsymbol{p}^{(i+1)} = \boldsymbol{r}^{(i+1)} + \beta_i \boldsymbol{p}^{(i)}, \end{cases} \qquad (7.5.18)$$

所以当$i = 0$时，

$$\begin{cases} \boldsymbol{r}^{(1)} = \boldsymbol{r}^{(0)} - \alpha_0 A\boldsymbol{p}^{(0)}, \\ \boldsymbol{p}^{(1)} = \boldsymbol{r}^{(1)} + \beta_0 \boldsymbol{p}^{(0)} = \boldsymbol{r}^{(1)} + \beta_0 \boldsymbol{r}^{(0)} \end{cases}$$

当$i = 1$时，

$$\begin{cases} \boldsymbol{r}^{(2)} = \boldsymbol{r}^{(1)} - \alpha_1 A\boldsymbol{p}^{(1)} = \boldsymbol{r}^{(0)} - \alpha_0 A\boldsymbol{p}^{(0)} - \alpha_1 A\boldsymbol{p}^{(1)}, \\ \boldsymbol{p}^{(2)} = \boldsymbol{r}^{(2)} + \beta_1 \boldsymbol{p}^{(1)} = \boldsymbol{r}^{(2)} + \beta_1 \boldsymbol{r}^{(1)} + \beta_1\beta_0 \boldsymbol{r}^{(0)}. \end{cases}$$

继续上述推导可得

$$r^{(i+1)} = r^{(0)} - \alpha_0 Ap^{(0)} - \alpha_1 Ap^{(1)} - \cdots - \alpha_i Ap^{(i)}, \tag{7.5.19}$$

$$p^{(i+1)} = r^{(i+1)} + \beta_i r^{(i)} + \beta_i \beta_{i-1} r^{(i-1)} + \cdots + \beta_i \beta_{i-1} \cdots \beta_0 r^{(0)}, \tag{7.5.20}$$

$$r^{(i+1)} = p^{(i+1)} - \beta_i p^{(i)}. \tag{7.5.21}$$

有了关系式(7.5.19)，(7.5.20)，(7.5.21)后，我们可以证明如下定理。

定理 7.5.3 剩余向量系$\{r^{(i)}\}$构成一个正交向量组，即

$$(r^{(i)}, r^{(j)}) = 0 \ (i \neq j), \tag{7.5.22}$$

方向向量系$\{p^{(i)}\}$构成一个A一共轭向量系，即

$$(p^{(i)}, Ap^{(j)}) = 0 \ (i \neq j). \tag{7.5.23}$$

证明 应用归纳法来证明。

$$(r^{(0)}, r^{(1)}) = (r^{(0)}, r^{(0)} - \alpha_0 Ap^{(0)}) = (r^{(0)}, r^{(0)}) - (r^{(0)}, \alpha_0 Ap^{(0)})$$

$$= (r^{(0)}, r^{(0)}) - \frac{(r^{(0)}, r^{(0)})}{(r^{(0)}, Ar^{(0)})}(r^{(0)}, Ar^{(0)}) = 0,$$

$$(p^{(0)}, Ap^{(1)}) = (r^{(0)}, A(r^{(1)} + \beta_0 r^{(0)})) = (r^{(0)}, Ar^{(1)})$$

$$+ \beta_0 (r^{(0)}, Ar^{(0)}) = (r^{(0)}, Ar^{(1)}) - \frac{(r^{(1)}, Ar^{(0)})}{(r^{(0)}, Ar^{(0)})}(r^{(0)}, Ar^{(0)})$$

$$= (r^{(0)}, Ar^{(1)}) - (r^{(1)}, Ar^{(0)}) = 0.$$

设$r^{(0)}, r^{(1)}, \cdots, r^{(k)}$已两两正交；$p^{(0)}, p^{(1)}, \cdots, p^{(k)}$已两两$A$一共轭，现在要证

$$(r^{(k+1)}, r^{(j)}) = 0 \quad (j = 0, 1, \cdots, k).$$

因为

$$\begin{aligned}(r^{(k+1)}, r^{(j)}) &= (r^{(k)} - \alpha_k Ap^{(k)}, r^{(j)}) \\ &= (r^{(k)}, r^{(j)}) - \alpha_k (Ap^{(k)}, r^{(j)}) \\ &= (r^{(k)}, r^{(j)}) - \alpha_k (Ap^{(k)}, p^{(j)} - \beta_{j-1} p^{(j-1)}) \\ &= (r^{(k)}, r^{(j)}) - \alpha_k (Ap^{(k)}, p^{(j)}) \\ &\quad + \alpha_k \beta_{j-1} (Ap^{(k)}, p^{(j-1)}) = 0,\end{aligned}$$

所以(7.5.22)成立。再证

$$(p^{(k+1)}, Ap^{(j)}) = 0 \quad (j=0, 1, \cdots, k).$$

因为

$$(p^{(k+1)}, Ap^{(j)}) = (r^{(k+1)} + \beta_k p^{(k)}, Ap^{(j)})$$
$$= (r^{(k+1)}, Ap^{(j)}) + \beta_k (p^{(k)}, Ap^{(j)}),$$

所以当$j=k$时，

$$(p^{(k+1)}, Ap^{(k)}) = (r^{(k+1)}, Ap^{(k)})$$
$$- \frac{(r^{(k+1)}, Ap^{(k)})}{(p^{(k)}, Ap^{(k)})} (p^{(k)}, Ap^{(k)}) = 0.$$

当$j \leqslant k-1$时，

$$(p^{(k+1)}, Ap^{(j)}) = (r^{(k+1)}, Ap^{(j)})$$
$$= (r^{(k+1)}, \frac{1}{\alpha_j}(r^{(j)} - r^{(j+1)}))$$
$$= \frac{1}{\alpha_j}[(r^{(k+1)}, r^{(j)}) - (r^{(k+1)}, r^{(j+1)})]$$
$$= 0,$$

从而证明了(7.5.23)。

下面将α_i和β_i的计算公式进行简化。由(7.5.20)和(7.5.22)可得

$$(r^{(i)}, p^{(i)}) = (r^{(i)}, r^{(i)} + \beta_{i-1} r^{(i-1)} + \cdots + \beta_{i-1} \cdots \beta_0 r^{(0)})$$
$$= (r^{(i)}, r^{(i)}),$$

于是，

$$\alpha_i = (r^{(i)}, p^{(i)}) / (p^{(i)}, Ap^{(i)}) = (r^{(i)}, r^{(i)}) / (p^{(i)}, Ap^{(i)}),$$

所以当$r^{(i)} \neq 0$时，$\alpha_i > 0$。又由(7.5.18)，(7.5.22)，(7.5.23)可得

$$(r^{(i)}, Ap^{(i-1)}) = (r^{(i)}, \frac{1}{\alpha_{i-1}}(r^{(i-1)} - r^{(i)}))$$
$$= -(r^{(i)}, r^{(i)}) / \alpha_{i-1},$$

$$(\boldsymbol{p}^{(i-1)}, A\boldsymbol{p}^{(i-1)}) = (\boldsymbol{r}^{(i-1)} + \beta_{i-2}\boldsymbol{p}^{(i-2)}, A\boldsymbol{p}^{(i-1)})$$
$$= (\boldsymbol{r}^{(i-1)}, A\boldsymbol{p}^{(i-1)})$$
$$= (\boldsymbol{r}^{(i-1)}, \frac{1}{\alpha_{i-1}}(\boldsymbol{r}^{(i-1)} - \boldsymbol{r}^{(i)}))$$
$$= (\boldsymbol{r}^{(i-1)}, \boldsymbol{r}^{(i-1)})/\alpha_{i-1},$$

于是

$$\beta_{i-1} = -\frac{(\boldsymbol{r}^{(i)}, A\boldsymbol{p}^{(i-1)})}{(\boldsymbol{p}^{(i-1)}, A\boldsymbol{p}^{(i-1)})} = \frac{(\boldsymbol{r}^{(i)}, \boldsymbol{r}^{(i)})}{(\boldsymbol{r}^{(i-1)}, \boldsymbol{r}^{(i-1)})}.$$

可见当 $\boldsymbol{r}^{(i)} \neq 0$ 时，$\beta_{i-1} > 0$，从而得到共轭斜量法的算法。

算法 7.5.1 设 A 为 n 阶对称正定阵，$\boldsymbol{b} = (b_1, b_2, \cdots, b_n)^T$；eps 为 $(\boldsymbol{r}, \boldsymbol{r})$ 或 $(A\boldsymbol{p}, \boldsymbol{p})$ 的允许误差，引入比例因子 d 以防发生溢出，取初始向量为 0 和限定迭代次数不超过 $3n$。

1) $d = 2^{\omega} \doteq \mathrm{Max}_i(|b_i|)$，其中 ω 为整数。

2) $\boldsymbol{p} = \boldsymbol{r} = \boldsymbol{b}/d$。

3) 对 $k = 0, 1, \cdots, 3n$，

 3.1) $q \leftarrow d \cdot (\boldsymbol{r}, \boldsymbol{r})/(A\boldsymbol{p}, \boldsymbol{p})$，

 3.2) 对 $i = 1, 2, \cdots, n$，

  $x_i \leftarrow x_i + qp_i$，

  $\bar{r}_i \leftarrow r_i - q\sum_{j=1}^{n} a_{ij}p_j$，$(\boldsymbol{r}, \boldsymbol{r}) \leqslant$ eps 转 6)，

 3.3) $e \leftarrow (\bar{\boldsymbol{r}}, \bar{\boldsymbol{r}})/(\boldsymbol{r}, \boldsymbol{r})$， $\boldsymbol{r} \leftarrow \bar{\boldsymbol{r}}$，

 3.4) $\boldsymbol{p} \leftarrow \boldsymbol{r} + e\boldsymbol{p}$，$(A\boldsymbol{p}, \boldsymbol{p}) \leqslant$ eps 转 6)，

 3.5) NEXT k。

4) $\bar{d} \leftarrow 2^{\omega} \doteq \max_i(|p_i|)$。

 5) $\boldsymbol{p} \leftarrow \boldsymbol{p}/\bar{d}$，$\boldsymbol{r} \leftarrow \boldsymbol{r}/\bar{d}$，$d \leftarrow \bar{d}\,d$ 转3)。

 6) 输出结果 $\boldsymbol{x}$ 和k。

例 7.5.1 用共轭斜量法解方程组

$$\begin{cases}3x_1 + x_2 = 5; \\ x_1 + 2x_2 = 5.\end{cases}$$

解 $A = \begin{bmatrix}3 & 1 \\ 1 & 2\end{bmatrix}$, $\quad b = \begin{bmatrix}5 \\ 5\end{bmatrix}$.

取

$$x^{(0)} = \begin{bmatrix}0 \\ 0\end{bmatrix},$$

1) 计算 $r^{(0)} = b - Ax^{(0)} = p^{(0)} = \begin{bmatrix}5 \\ 5\end{bmatrix}$,

$$Ap^{(0)} = \begin{bmatrix}20 \\ 15\end{bmatrix}, \quad (p^{(0)}, \ Ap^{(0)}) = 175,$$

$$(r^{(0)}, r^{(0)}) = 50, \quad \alpha_0 = 2/7.$$

得

$$x^{(1)} = x^{(0)} + \alpha_0 p^{(0)} = \frac{2}{7}\begin{bmatrix}5 \\ 5\end{bmatrix}.$$

2) $r^{(1)} = \frac{5}{7}\begin{bmatrix}-1 \\ 1\end{bmatrix}$, $\beta_0 = 1/49$,

$$p^{(1)} = r^{(1)} + \beta_0 p^{(0)} = \frac{1}{49}\begin{bmatrix}-30 \\ 40\end{bmatrix},$$

$$Ap^{(1)} = \frac{1}{49}\begin{bmatrix}-50 \\ 50\end{bmatrix}, (r^{(1)}, r^{(1)}) = \frac{50}{49}.$$

$$(p^{(1)}, \ Ap^{(1)}) = \frac{3500}{49^2}, \ \alpha_1 = \frac{7}{10},$$

得

$$x^{(2)} = x^{(1)} + \alpha_1 p^{(1)} = \begin{bmatrix}1 \\ 2\end{bmatrix} = x^*,$$

这时

$$r^{(2)} = \begin{bmatrix}0 \\ 0\end{bmatrix}.$$

下面我们将证明，用共轭斜量法所求得 $x^{(i)}$，随着 i 的增大而一步步地逼近方程组的精确解，而且若在计算中没有误差

时，至多到 n 步便能得到方程组的精确解。

定理 7.5.4 依共轭斜量法求解 n 阶线性方程组，最多计算 n 步便可得到方程组（7.5.1）的精确解。

证明 根据定理 7.5.3 可知，如果 $n+1$ 个剩余向量 $\boldsymbol{r}^{(0)}$, $\boldsymbol{r}^{(1)}$, …, $\boldsymbol{r}^{(n)}$ 均不为零，那末这些向量必线性无关，从而与线性无关向量组的性质发生矛盾，故在 $\boldsymbol{r}^{(0)},\boldsymbol{r}^{(1)}$, …, $\boldsymbol{r}^{(n)}$ 中至少有一个 $\boldsymbol{r}^{(k)}$ 为零，由 $\boldsymbol{r}^{(k)}=\boldsymbol{b}-A\boldsymbol{x}^{(k)}=0$，得 $\boldsymbol{x}^{(k)}=\boldsymbol{x}^*$.

定理 7.5.5 设当 $l<k$ 时，$\boldsymbol{x}^{(k)}$ 比 $\boldsymbol{x}^{(l)}$ 更接近于精确解 $\boldsymbol{x}^*$. 则

$$\|\boldsymbol{x}^{(l)}-\boldsymbol{x}^*\|\geqslant\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|.$$

证明 如果能证明

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|\geqslant\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|, \tag{7.5.24}$$

那末，用不等式进行递推，便能得到定理7.5.5的结论.

因为 $\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\alpha_k\boldsymbol{p}^{(k)}$，所以

$$\begin{aligned}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\|^2&=(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*-\alpha_k\boldsymbol{p}^{(k)},\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*-\alpha_k\boldsymbol{p}^{(k)})\\&=\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|^2-2\alpha_k(\boldsymbol{p}^{(k)},\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*)+\alpha_k^2(\boldsymbol{p}^{(k)},\boldsymbol{p}^{(k)}).\end{aligned} \tag{7.5.25}$$

由于 $\alpha_k\geqslant0$，$(\boldsymbol{p}^{(k)},\boldsymbol{p}^{(k)})>0$，所以只要证明 $(\boldsymbol{p}^{(k)},\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*)\leqslant0$，定理的结论就成立了. 因为

$$\begin{aligned}\boldsymbol{x}^{(k+1)}&=\boldsymbol{x}^{(k)}+\alpha_k\boldsymbol{p}^{(k)}=\boldsymbol{x}^{(k-1)}+\alpha_{k-1}\boldsymbol{p}^{(k-1)}+\alpha_k\boldsymbol{p}^{(k)}=\cdots\\&=\boldsymbol{x}^{(0)}+\alpha_0\boldsymbol{p}^{(0)}+\alpha_1\boldsymbol{p}^{(1)}+\cdots+\alpha_k\boldsymbol{p}^{(k)},\end{aligned}$$

据定理 7.5.4 必有 $\boldsymbol{x}^{(m)}=\boldsymbol{x}^*$ $(m\leqslant n)$，所以 $\boldsymbol{x}^{(m)}=\boldsymbol{x}^*=\boldsymbol{x}^{(0)}+\alpha_0\boldsymbol{p}^{(0)}+\cdots+\alpha_{m-1}\boldsymbol{p}^{(m-1)}$. 因为当 $k=m-1$ 时，(7.5.24) 显然成立. 于是只需讨论 $k<m-1$ 的情况. 此时，由

$$\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*=-\alpha_{k+1}\boldsymbol{p}^{(k+1)}-\alpha_{k+2}\boldsymbol{p}^{(k+2)}-\cdots-\alpha_{m-1}\boldsymbol{p}^{(m-1)},$$

则

$$\begin{aligned}(\boldsymbol{p}^{(k)},\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*)=&-\alpha_{k+1}(\boldsymbol{p}^{(k)},\boldsymbol{p}^{(k+1)})-\alpha_{k+2}(\boldsymbol{p}^{(k)},\boldsymbol{p}^{(k+2)})\\&-\cdots-\alpha_{m-1}(\boldsymbol{p}^{(k)},\boldsymbol{p}^{(m-1)}).\end{aligned}$$

又由于$\alpha_{k+i}\geqslant 0$，而且容易证明

$$(p^{(k)},p^{(k+1)})=\|r^{(k)}\|^2\|r^{(k+1)}\|^2\sum_{j=0}^{k}(1/\|r^{(j)}\|^2)\geqslant 0, \tag{7.5.26}$$

从而$(p^{(k)},x^{(k+1)}-x^*)\leqslant 0$，即本定理成立。

共轭斜量法虽然本质上是一种直接法，在有限步内可以得到方程组的精确解。可是，在实际计算中由于存在舍入误差，使剩余向量不能精确满足正交关系，所以可能有$r^{(n)}\neq 0$，而且当系数矩阵病态时，$r^{(n)}$偏离零的程度也更厉害。不过，由于共轭斜量法的计算公式具有迭代法的特点，如果$r^{(n)}\neq 0$，还可以用此法继续计算下去，只要$r^{(n+i)}\neq 0$，$F(x)$的值总会减小，直到由于舍入误差的影响使近似解不能再改进为止。因此，最后总可以得到更好一些的近似解。

本章讨论了解线性方程组的线性与非线性迭代法，重点介绍了 SOR 法，Chebyshev 半迭代法和共轭斜量法，这三种方法也是常用的方法。

Jacobi 法与 Gauss-Siedel 法的收敛性并不互相包含，不过当两者皆收敛时，往往后者收敛快。当收敛慢时，可用 Chebyshev 加速，从而得到 Chebyshev 半迭代法。当原迭代矩阵B的特征值全部是实数时，Chebyshev 加速是很有效的。但是当原方程的系数矩阵A是具有相容次序的对称正定阵时，SOR方法的收敛速度比简单迭代法的 Chebyshev 加速还要快两倍。即使矩阵不具有相容次序，在某些条件下也还比 Jacobi 方法或 Gauss-Siedel 方法快一个数量级，因此，在实用中常常使用 SOR 方法，其松弛因子的选择往往依赖于实际经验。

当系数矩阵A是对称正定阵时，共轭斜量法是一种较好的方法。它不需要估计参数，但是存在着计算过程中舍入误差比较敏感的缺点。近年来提出的“不完全分解预处理共轭斜量

法”，弥补了这方面的缺点。这种方法虽然在理论上还不成熟，但是在某些实际应用中，却取得了惊人的效果，受到了国内外的普遍重视。

近年来，对迭代法还常常用预处理法的技巧来减小原矩阵的条件数以加快迭代法的收敛速度。尤其对病态方程组效果特别显著。另外，还发展了不少对某些具体问题的解决特别有效的迭代法，如“强隐式迭代法”等。最后还应指出，迭代法的使用与具体问题的特点有着很密切的联系。在各种实际问题中使用迭代法的经验与针对问题进行详细讨论是很重要的。

第七章习题

7.1 设线性方程组 $Ax=b$ 中，A 为

$$1)\quad \begin{pmatrix} 3 & 1 & 2 \\ 0 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix};\qquad 2)\quad \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix},$$

若用 Jacobi 迭代法，Gauss-Siedel 迭代法求解，是否收敛？若收敛，求方程组的解。

7.2 设 $Ax=b$，其中 $A\in R^{n\times n}$，试证明

$$\max_i \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1$$

是 Jacobi 迭代法收敛的充分条件。

7.3 用逐次超松弛迭代法解方程组（取 $\omega=0.9$）

$$\begin{cases} 5x_1+2x_2+x_3=-12, \\ -x_1+4x_2+2x_3=20, \\ 2x_1-3x_2+10x_3=3 \end{cases}$$

准确到小数点后4位。

7.4 证明定理7.2.3和定理7.2.5的充分性。

7.5 试证明 n 阶矩阵 A 不可约的充要条件是

$n=1$或者对任意两个不同的整数 i，j，其中$1\leqslant i\leqslant n$，$1\leqslant j\leqslant n$，则 $a_{ij}\neq 0$ 或者存在i_1，i_2，…，i_s满足

$$a_{ii_1}\cdot a_{i_1i_2}\cdot\cdots\cdot a_{i_sj}\neq 0.$$

7.6　证明 $\lim\limits_{m\to\infty}(\|B^m\|)^{1/m}=S(B)$。

7.7　假若A为正定矩阵，那末使用松弛因子 $\omega_1,\omega_2,\cdots$的SOR方法收敛，下列条件至少一个成立

1)　对某一$\varepsilon>0$，对足够大的 i，有

$$\varepsilon\leqslant\omega_i\leqslant 2-\varepsilon\quad i=1,2,\cdots.$$

2)　对所有足够大的 i，$0\leqslant\omega_i\leqslant 2$ 和级数

$$\sum_{i=1}^{\infty}\omega_i(2-\omega_i)$$

发散。

7.8　当 $\omega=1$时，对（7.3.12）使用简单迭代法，证明该迭代法收敛。

7.9　设序列 $\{x_i\}=\{(-1)^{i+1}\}$，证明序列 $\{x_i\}$ 为 Cesàro 可求和。问是否存在 $\sum\limits_{k=0}^{n}a_{nk}$ $n=0,1,2,\cdots$，使作成的新序列 $y_n=\sum\limits_{k=0}^{n}a_nkxk$ 发散。

7.10　设方程组 $x=Bx+k$ 与$Ax=b$（其中 A 为非奇异矩阵）等价，且矩阵 B 的特征值 μ 满足：$-2\leqslant\mu\leqslant 0.9$，$1.2\leqslant\mu\leqslant 3$，使用线性定常迭代法

$$x^{(n+1)}=Bx^{(n)}+k,$$

构造一个收敛的半迭代法。

7.11　对习题7.3 中的方程组分别利用基于 Jacobi 和 Gauss-Siedel 迭代法的 Chebyshev 迭代法求解。

7.12　证明公式（7.5.26）成立。

7.13　使用共轭斜量法解方程组

$$\begin{cases}4x_1-x_2+2x_3=12,\\-x_1+5x_2+3x_3=10,\\2x_1+3x_2+6x_3=18.\end{cases}$$

7.14　设 $Ax=b$ 其中 A 为非奇异阵，求证 A^TA 为对称正定阵，并写出用共轭斜量法解 $A^TAx=A^Tb$ 的计算公式。

7.15　设 A 为对称正定阵，求证 $Ax=b$ 的另一种共轭斜量法的迭代公式

为

$$
\begin{cases}
x^{(0)}\text{初始向量，} \Delta x_{-1} = \Delta r_{-1} = 0, \\
x^{(k+1)} = x^{(k)} + \Delta x_k, \\
\Delta x_k = \dfrac{1}{q_k}(r^{(k)} + e_k \Delta x_{k-1}), \\
q_k = \dfrac{(r^{(k)}, Ar^{(k)})}{(r^{(k)}, r^{(k)})} - e_k, \quad e_0 = 0, \\
e_k = \dfrac{(r^{(k)}, r^{(k)})}{(r^{(k-1)}, r^{(k-1)})} q_{k-1}, \\
r^{(k+1)} = r^{(k)} + \Delta r_k, \\
\Delta r_{k+1} = \dfrac{1}{q_k}(-Ar^{(k)} + e_k \Delta r_{k-1}). \\
k = 0, 1, 2, \cdots.
\end{cases}
$$

参　考　书

[1] 徐树荣译，线性代数方程组的计算机解法，科学出版社。

[2] D. M. Young. "*Iterative Solution of large linear systems*", Academic press, New York and London, 1971.

[3] Stewart, G. "*Introduction to Matrix Computation*" Academic press. New York 1973.

[4] Jennings, A. "*Matrix computation for Engineers and Scientists*", London New york. Sydney. Toronto, 1977.

[5] 冯康等编，数值计算方法，国防工业出版社。1978.

[6] 曹志浩等编。矩阵计算和方程求根。1979.

第八章　非线性方程组的牛顿迭代解法

§1　引　　言

近年来，许多应用领域和数值分析本身都提出了多个实变量的非线性方程组

$$f_i(\xi_1, \xi_2, \cdots, \xi_n) = 0 \qquad (i = 1, \cdots, n) \tag{8.1.1}$$

的数值求解问题。因此研究这一课题显得十分迫切。

非线性方程组（8.1.1）也可用向量形式表示：

$$\boldsymbol{F}(\boldsymbol{x}) = 0$$

这里$\boldsymbol{F} = (f_1, f_2, \cdots, f_n)^T$，$\boldsymbol{x} = (\xi_1, \xi_2, \cdots, \xi_n)^T$。向量函数$\boldsymbol{F}(\boldsymbol{x})$可视为一个从区域 $D \subset \boldsymbol{R}^n$ 至$\boldsymbol{R}^n$ 的映射，常记作 $\boldsymbol{F}$：$D \subset \boldsymbol{R}^n \to \boldsymbol{R}^n$。如果 f_1，f_2，…，f_n 皆为线性函数，则（8.1.1）即为 n 个变量的线性方程组；如果 $n = 1$，则(8.1.1)即为单变量的非线性方程。这两个特例在数值分析教程中均作了较详细的研究。对于一般的非线性方程组的数值求解，很自然地会面临两个问题：

(1) 方程组（8.1.1）的可解性分析；

(2) 寻找有效的数值方法来求解（8.1.1）。

所谓有效的意思，主要是指迭代算法具有较快的收敛速度和较少的运算量。第一个问题已远远超出本教程的范围，因为它属于非线性泛函分析的范畴。所以，以下我们均假定方程组（8.1.1）存在解。今后的任务仅在于构造各种迭代算法，以求得（8.1.1）的近似解。

对于许多迭代算法常成立以下定理：如果 x^* 为（8.1.1）在区域 D 中的一个解，而 x_0 为一个充分接近 $x^* \in D$ 的初始近似解，则由 x_0 出发产生的迭代序列 $\{x_k\}$ 恒有定义且收敛于 x^*. 这种类型的定理常称为局部收敛定理. 固然，有极个别的情形，迭代序列 $\{x_k\}$ 收敛于 x^* 与初始近似解 x_0 的选取无关. 这种收敛性质常称为整体收敛性. 对于某迭代算法，若 n 维空间中存在球 $S(x^*, r) \subset D$，从该球中任何一点出发，由此生成的 $\{x_k\}$ 必收敛于 x^*，则球 S 就称为该迭代算法的局部收敛域. 非线性方程组的迭代解法的基本思想由以下三部分组成：

(1) 由一个极差的初始估猜 x_0 出发，设法进入局部收敛区域. 在实际问题中，有时事先对所求的解已经有了一个较好的估计，则这一步就可以免去.

(2) 从局部收敛区域 S 中一个初始近似解 x_0 出发，通过加速收敛的迭代算法，求出足够精确的近似解. 以后着重介绍的就是加速收敛的局部技巧，或者说局部方法.

(3) 继续求方程组的其余解.

§2 非线性方程组的牛顿迭代法

正如一维情形那样，牛顿法也是求解非线性方程组（8.1.1）的一个基本算法. 许多其它的迭代算法都是在它的基础上作各种修改得到的. 因此，无论在实践和理论方面它都具有重要意义.

一维牛顿法的几何意义是：在横坐标为 x_k 的点 P_k 引切线. 于是在 P_k 点近傍以该切线近似替代曲线 $y=f(x)$. 然后，将该切线与 x 轴的交点的横坐标 x_{k+1} 作为根 x^* 的新近似值. 如此不断迭代，直至得到满意的近似值为止. 现在，我们就来仿

照一维情形，构造出非线性方程组对应的牛顿迭代公式。

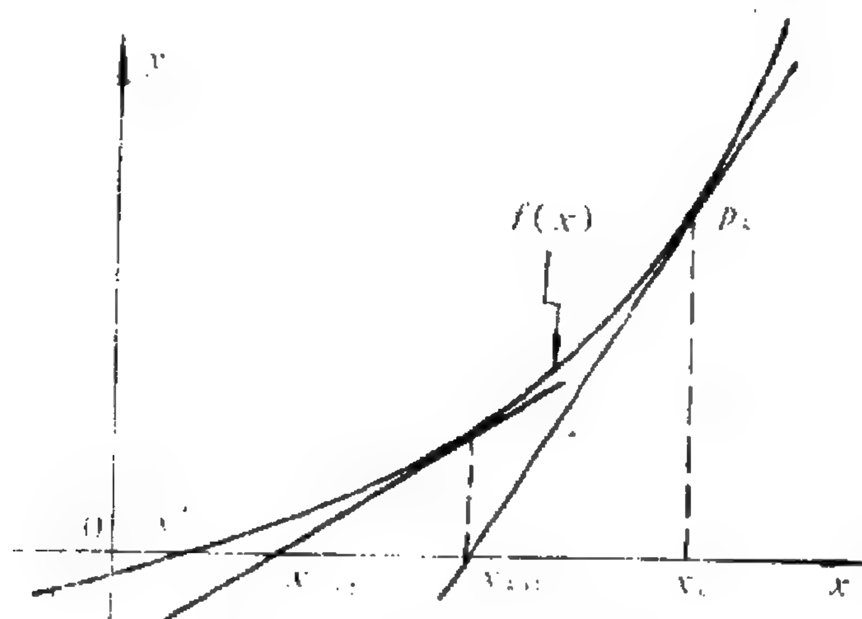


图8.1　牛顿法的几何意义

为叙述方便起见，仅对二个方程的方程组

$$
\begin{cases} f_1(\xi_1, \xi_2) = 0, \\ f_2(\xi_1, \xi_2) = 0 \end{cases} \tag{8.2.1}
$$

来讨论。假定在方程组（8.2.1）解 x^* 的某邻域内，f_1、f_2 连续且有连续的一阶偏导数。现设（$\xi_1^{(k)}$，$\xi_2^{(k)}$）为上述邻域内的点，作线性函数

$$
\begin{cases} l_1(\xi_1, \xi_2) = f_1(\xi_1^{(k)}, \xi_2^{(k)}) + \left.\dfrac{\partial f_1}{\partial \xi_1}\right|_{(\xi_1^{(k)}, \xi_2^{(k)})}(\xi_1 - \xi_1^{(k)}) \\ \qquad + \left.\dfrac{\partial f_1}{\partial \xi_2}\right|_{(\xi_1^{(k)}, \xi_2^{(k)})}(\xi_2 - \xi_2^{(k)}), \\ l_2(\xi_1, \xi_2) = f_2(\xi_1^{(k)}, \xi_2^{(k)}) + \left.\dfrac{\partial f_2}{\partial \xi_1}\right|_{(\xi_1^{(k)}, \xi_2^{(k)})}(\xi_1 - \xi_1^{(k)}) \\ \qquad + \left.\dfrac{\partial f_2}{\partial \xi_2}\right|_{(\xi_1^{(k)}, \xi_2^{(k)})}(\xi_2 - \xi_2^{(k)}), \end{cases} \tag{8.2.2}
$$

在该邻域内近似替代函数 $f_1(\xi_1, \xi_2)$ 和 $f_2(\xi_1, \xi_2)$。于是，非线性方程组（8.2.1）即以线性方程组

$$\begin{cases} \left.\dfrac{\partial f_1}{\partial \xi_1}\right|_{x=x_k} \Delta\xi_1^{(k)} + \left.\dfrac{\partial f_1}{\partial \xi_2}\right|_{x=x_k} \Delta\xi_2^{(k)} = -f_1(\boldsymbol{x}_k), \\ \left.\dfrac{\partial f_2}{\partial \xi_1}\right|_{x=x_k} \Delta\xi_1^{(k)} + \left.\dfrac{\partial f_2}{\partial \xi_2}\right|_{x=x_k} \Delta\xi_2^{(k)} = -f_2(\boldsymbol{x}_k) \end{cases} \tag{8.2.3}$$

近似替代，其中 $\Delta\xi_1^{(k)} = \xi_1 - \xi_1^{(k)}$，$\Delta\xi_2^{(k)} = \xi_2 - \xi_2^{(k)}$，$\boldsymbol{x}_k = (\xi_1^{(k)}, \xi_2^{(k)})^T$。如果在上述邻域中矩阵

$$\begin{pmatrix} \dfrac{\partial f_1}{\partial \xi_1} & \dfrac{\partial f_1}{\partial \xi_2} \\ \dfrac{\partial f_2}{\partial \xi_1} & \dfrac{\partial f_2}{\partial \xi_2} \end{pmatrix} \tag{8.2.4}$$

为非异，则可解得

$$\begin{pmatrix} \Delta\xi_1^{(k)} \\ \Delta\xi_2^{(k)} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial f_1}{\partial \xi_1} & \dfrac{\partial f_1}{\partial \xi_2} \\ \dfrac{\partial f_2}{\partial \xi_1} & \dfrac{\partial f_2}{\partial \xi_2} \end{pmatrix}_{x=x_k}^{-1} \begin{pmatrix} -f_1 \\ -f_2 \end{pmatrix}_{x=\boldsymbol{x}_k}. \tag{8.2.5}$$

矩阵 (8.2.4) 常称为向量函数 $\boldsymbol{F}(\boldsymbol{x})$ 的 Jacobi 矩阵，记作 $\boldsymbol{F}'(\boldsymbol{x})$。进而，记 $\xi_1^{(k+1)} = \xi_1^{(k)} + \Delta\xi_1^{(k)}$，$\xi_2^{(k+1)} = \xi_2^{(k)} + \Delta\xi_2^{(k)}$，则 (8.2.5) 即可写作

$$\Delta\boldsymbol{x}_k = -\boldsymbol{F}'(\boldsymbol{x}_k)^{-1}\boldsymbol{F}(\boldsymbol{x}_k)$$

或

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{F}'(\boldsymbol{x}_k)^{-1}\boldsymbol{F}(\boldsymbol{x}_k). \tag{8.2.6}$$

公式 (8.2.6) 称为求解非线性方程组 (8.2.1) 的牛顿迭代公式，而线性方程组 (8.2.3) 则称为牛顿方程组。

我们对向量函数 $\boldsymbol{F}(\boldsymbol{x})$ 自始至终作如下假设：

假设 1.

(1) 向量函数 $\boldsymbol{F}(\boldsymbol{x})$ 在开的凸集 D*) 内连续可微，即在 D 内函数 $f_i(\xi_1, \xi_2, \cdots, \xi_n)$ $(i=1, \cdots, n)$ 连续且有连续的一阶偏导数（请注意：下文中集合 D 总是假定为开的凸集，以后不再说明）。

(2) D内存在点x^*，使$F(x^*)=0$且$F'(x^*)$为非异，$F'(x)$为 Jacobi 矩阵

$$\begin{pmatrix} \frac{\partial f_1}{\partial \xi_1} & \frac{\partial f_1}{\partial \xi_2} & \cdots & \frac{\partial f_1}{\partial \xi_n} \\ \frac{\partial f_2}{\partial \xi_1} & \frac{\partial f_2}{\partial \xi_2} & \cdots & \frac{\partial f_2}{\partial \xi_n} \\ \cdots & & \cdots & \cdots \\ \frac{\partial f_n}{\partial \xi_1} & \frac{\partial f_n}{\partial \xi_2} & \cdots & \frac{\partial f_n}{\partial \xi_n} \end{pmatrix}. \tag{8.2.7}$$

综上所述，牛顿法的基本思想是，将非线性方程组逐次线性化，从而形成迭代算法。具体地说，设x_k为x^*的第k次近似，在x_k的某邻域中，如果线性函数

$$L_k(x)=F(x_k)+F'(x_k)(x-x_k) \tag{8.2.8}$$

是向量函数$F(x)$的一个较好的逼近，则由线性方程组$L_k(x)=0$求得的解x_{k+1}必然是x^*更好的近似解。

牛顿法的计算步骤可总结成如下算法。

算法8.2.1 设$F: R^n \to R^n$，x_0为x^*的初始近似，它们都满足定理8.4.1或定理8.4.2的条件，欲求满足一定精度的近似解。

1) 对$k=0, 1, \cdots, m$(m为允许的最大迭代次数)，

  1.1) 计算$F(x_k)$，如果x_k满足精度要求，则转 3)。

  1.2) 计算$F'(x_k)$。

  1.3) 求解线性方程组$F'(x_k)\Delta x_k=-F(x_k)$，置$x_{k+1}=x_k+\Delta x_k$。

2) NEXT k。

3) 置$x^*=x_{k+1}$。

4) END。

牛顿法的特点是，把向量函数$F(x)$看成一个整体，通过

线性化，从而形成迭代算法，这种方法的优点是，便于收敛性分析，利于迭代格式的修正，但这种方法并未充分利用函数的具体结构。对于一个非线性向量函数 $\boldsymbol{F}(\boldsymbol{x})$，其非线性程度在各个分量函数 $f_i(\boldsymbol{x})$ $(i=1,\cdots,n)$ 上分布可能是不均衡的。也就是说，有的分量非线性程度高；有的则接近线性函数，甚至有的完全是线性函数。在这种情况下，把所有分量 $f_i(\boldsymbol{x})$ 采取完全相同的数值处理，将不利于方法整体计算效率的提高。于是人们设想：采取对 f_i $(i=1,\cdots,n)$ 逐个线性化结合代入法以形成迭代算法的途径，这也就是 **Brown** 算法的基本思想。下面简要叙述 **Brown** 方法的迭代步骤。

设 $\boldsymbol{x}_k$ 为 $\boldsymbol{x}^*$ 的第 k 次近似。首先将 $f_1(\boldsymbol{x})$ 在 $\boldsymbol{x}_k$ 近傍线性化，当 $\boldsymbol{x}_k$ 充分接近 $\boldsymbol{x}^*$ 时，即有

$$f_1(\boldsymbol{x}) \approx f_1(\boldsymbol{x}_k) + \sum_{j=1}^{n} \frac{\partial f_1(\boldsymbol{x}_k)}{\partial \xi_j}(\xi_j - \xi_j^{(k)}).$$

令上式为零，则可解出一个变量，设为 ξ_n。则

$$\xi_n = \xi_n^{(k)} - \left(\frac{\partial f_1(\boldsymbol{x}_k)}{\partial \xi_n}\right)^{-1}\left\{\sum_{j=1}^{n-1} \frac{\partial f_1(\boldsymbol{x}_k)}{\partial \xi_j}(\xi_j - \xi_j^{(k)}) + f_1(\boldsymbol{x}_k)\right\}. \tag{8.2.9}$$

上式右端是变量 $\xi_1, \xi_2, \cdots, \xi_{n-1}$ 的线性函数。定义

$$L_n(\xi_1, \xi_2, \cdots, \xi_{n-1}) = \xi_n^{(k)} - \left(\frac{\partial f_1(\boldsymbol{x}_k)}{\partial \xi_n}\right)^{-1} \times \left\{\sum_{j=1}^{n-1} \frac{\partial f_1(\boldsymbol{x}_k)}{\partial \xi_j}(\xi_j - \xi_j^{(k)}) + f_1(\boldsymbol{x}_k)\right\}. \tag{8.2.10}$$

再考虑第二个分量函数 $f_2(\boldsymbol{x})$，把 $L_n(\xi_1, \xi_2, \cdots, \xi_{n-1})$ 作为函数变量 ξ_n 的近似，把它代入 $f_2(\boldsymbol{x})$，这样 $f_2(\boldsymbol{x})$ 就变成一个 $n-1$ 个变量的多元函数，即

$$g_2(\boldsymbol{x}) \equiv f_2(\xi_1, \xi_2, \cdots, \xi_{n-1}, L(\xi_1, \cdots, \xi_{n-1})) \tag{8.2.11}$$

对 $g_2(\boldsymbol{x})$ 作类似于 $f_1(\boldsymbol{x})$ 的处理，即能解得

$$\xi_{n-1}=\xi_{n-1}^{(k)}-\left(\frac{\partial g_2(\boldsymbol{x}_k)}{\partial \xi_{n-1}}\right)^{-1}\left\{\sum_{j=1}^{n-2}\frac{\partial g_2(\boldsymbol{x}_k)}{\partial \xi_j}(\xi_j-\xi_j^{(k)})\right.$$

$$\left.+g_2(\boldsymbol{x}_k)\right\}. \qquad (8.2.12)$$

注意上式中 $\boldsymbol{x}_k=(\xi_1^{(k)},\cdots,\ \xi_{n-1}^{(k)})^T$。上式右端为变量 $\xi_1,\xi_2,\cdots,\xi_{n-2}$ 的线性函数，记作 $L_{n-1}(\xi_1,\cdots,\ \xi_{n-2})$。

按照这样的步骤，依次处理 $\boldsymbol{F}(\boldsymbol{x})$ 的各个分量函数。最后考虑 $g_n(\boldsymbol{x})$，此时 $g_n(\boldsymbol{x})$ 已简化为只含 ξ_1 的单变量函数。通过线性化，便能解得

$$\xi_1=\xi_1^{(k)}-\left(\frac{dg_n(\boldsymbol{x}_k)}{d\xi_1}\right)^{-1}g_n(\boldsymbol{x}_k), \qquad (8.2.13)$$

于是即求得了 ξ_1^* 的第 $k+1$ 个近似值。由此出发利用线性函数 $L_2,\ L_3,\cdots,\ L_n$ 逐一回代，最后将能求得新的迭代点 $\boldsymbol{x}_{k+1}$。现举例说明以上迭代过程。

设非线性方程组为

$$\begin{cases} f_1(\boldsymbol{x})\equiv\xi_1\xi_3-2\xi_1+1=0, \\ f_2(\boldsymbol{x})\equiv\xi_1+5\xi_2+1=0, \\ f_3(\boldsymbol{x})\equiv\xi_1^2+\xi_2^2-\xi_3-2=0; \end{cases}$$

初始近似解为 $\boldsymbol{x}_0=(-2,\ 0,\ 2)^T$。

根据前述步骤，首先在点 $\boldsymbol{x}_k$ 近傍用线性函数

$$f_1(\boldsymbol{x}_k)+(\xi_1-\xi_1^{(k)})(\xi_3^{(k)}-2)+(\xi_3-\xi_3^{(k)})\xi_1^{(k)}$$

来近似替代 $f_1(\boldsymbol{x})$。令上式为零，即可解出变量

$$\xi_3=\xi_3^{(k)}-(\xi_1^{(k)})^{-1}[(\xi_3^{(k)}-2)\xi_1+1].$$

定义

$$L_3(\xi_1,\ \xi_2)=\xi_3^{(k)}-(\xi_1^{(k)})^{-1}[(\xi_3^{(k)}-2)\xi_1+1].$$

下一步即可利用 $L_3(\xi_1,\ \xi_2)$ 消去 $f_2(\boldsymbol{x})$ 中的变量 ξ_3（当然在本问题中，$f_2(\boldsymbol{x})$ 中并不含 ξ_3），从而定义新函数

$$g_2(\xi_1,\ \xi_2)\equiv\xi_1+5\xi_2+1.$$

$g_2(\xi_1,\ \xi_2)$ 在 $(\xi_1^{(k)},\ \xi_2^{(k)})$ 处的线性近似表达式为

$$g_2(\xi_1^{(k)},\ \xi_2^{(k)}) + (\xi_1 - \xi_1^{(k)}) + 5(\xi_2 - \xi_2^{(k)}).$$

令上式为零，即可解出变量

$$\xi_2 = -\frac{1}{5}(1+\xi_1).$$

定义

$$L_2(\xi_1) \equiv -\frac{1}{5}(1+\xi_1)$$

利用 $L_3(\xi_1, \xi_2)$ 和 $L_2(\xi_1)$ 消去 $f_3(x)$ 中的变量 ξ_2，ξ_3，从而定义新函数

$$g_3(\xi_1) \equiv \xi_1^2 + \frac{1}{25}(1+\xi_1)^2$$

$$-\{\xi_3^{(k)} - (\xi_1^{(k)})^{-1}[(\xi_3^{(k)}-2)\xi_1 + 1]\} - 2$$

$g_3(\xi_1)$ 在点 $\xi_1^{(k)}$ 近傍用线性函数

$$g_3(\xi_1^{(k)}) + \left(\frac{52}{25}\xi_1^{(k)} + \frac{\xi_3^{(k)} - 2}{\xi_1^{(k)}} + \frac{2}{25}\right)(\xi_1 - \xi_1^{(k)})$$

近似替代。令上式为零，即可解得

$$\xi_1^{k+1} = \xi_1^{(k)} -$$

$$\frac{(\xi_1^{(k)})^2 + \frac{1}{25}(1+\xi_1^{(k)})^2 - \{\xi_3^{(k)} - (\xi_1^{(k)})^{-1}[(\xi_3^{(k)}-2)\xi_1^{(k)} + 1] - 2}{\frac{52}{25}\xi_1^{(k)} + \frac{\xi_3^{(k)} - 2}{\xi_1^{(k)}} + \frac{2}{25}}$$

利用以上公式从 x_0 出发逐次迭代，可得到如下结果：

n	ξ_1	ξ_2	ξ_3
0	−2.000000	0.000000	2.000000
1	−2.112745	0.222549	2.500000
2	−2.103978	0.220796	2.475393
3	−2.103937	0.220787	2.475299
4	−2.103937	0.220787	2.475299

综上所述，Brown 算法是对非线性方程组逐个线性化，逐个消去变量以及向后回代相结合的迭代过程。该算法在§5还将详细讨论。

在结束本节之前，还有一个问题须作出回答。前面我们将向量函数 $\boldsymbol{F}(\boldsymbol{x})$ 的 Jacobi 矩阵用导数记号 $\boldsymbol{F}'(\boldsymbol{x})$ 来表示，那末，$\boldsymbol{F}(\boldsymbol{x})$ 的 Jacobi 矩阵究竟如何体现出"导数"的意义呢？

回想单变量实值函数在点 x 处导数的概念。如果存在一个实数 $a=f'(x)$，使得

$$\lim_{h\to 0}\left(\frac{1}{h}\right)[f(x+h)-f(x)-ah]=0$$

成立，则 a 即为函数 $f(x)$ 在点 x 处的导数。这个定义可以自然地推广到 n 维。

定义 8.2.1 设 $\boldsymbol{F}$: $D\subset \boldsymbol{R}^n\to\boldsymbol{R}^m$，$\boldsymbol{x}$ 为 D 的内点。若存在 $A\in\boldsymbol{R}^{m\times n}$ 使对所有 $\boldsymbol{h}\in\boldsymbol{R}^n$ 有

$$\lim_{h\to 0}\left(\frac{1}{\|\boldsymbol{h}\|}\right)[\|\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h})-\boldsymbol{F}(\boldsymbol{x})-A\boldsymbol{h}\|]=0 \qquad (8.2.14)$$

或

$$\boldsymbol{F}(\boldsymbol{x}+\boldsymbol{h})=\boldsymbol{F}(\boldsymbol{x})+A\boldsymbol{h}+o(\|\boldsymbol{h}\|). \qquad (8.2.15)$$

成立，则称 $\boldsymbol{F}(\boldsymbol{x})$ 在点 $\boldsymbol{x}$ 处为 Fréchet 可微（或 F-可微）。

据范数等价定理知，(8.2.14) 中的极限与 $\boldsymbol{R}^m$ 上所取的范数无关。也就是说，在某种范数意义下，如果 $\boldsymbol{F}(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处为 $\boldsymbol{F}$-可微，则对任何范数而言，$\boldsymbol{F}(\boldsymbol{x})$ 在 $\boldsymbol{x}$ 处仍为 $\boldsymbol{F}$-可微。不难证明：使 (8.2.14) 成立的矩阵 A 至多只有一个。

现设法用 $\boldsymbol{F}$ 的分量函数 $f_1,\cdots,f_m$ 的偏导数来表示矩阵 A。设 $A=(a_{i,j})$，因为极限式 (8.2.14) 对任何趋于零的向量 $\boldsymbol{h}\in\boldsymbol{R}^n$ 均成立，故若取 $\boldsymbol{h}=t\boldsymbol{e}_j\,(t>0)$，则

$$\lim_{t\to 0}\left(\frac{1}{t}\right)[f_i(\boldsymbol{x}+t\boldsymbol{e}_j)-f_i(\boldsymbol{x})-ta_{ij}]=0$$

$$(i=1,\cdots,m).$$

据假定，$f_i(i=1,\cdots,m)$ 在 x 处的偏导数存在，故上式意味着

$$a_{i,j}=\frac{\partial f_i}{\partial \xi_j}\quad (i=1,\cdots,m;\ j=1,\cdots,n).$$

由此可知，在极限式（8.2.14）成立的前提下，$F(x)$ 的Jacobi矩阵就是（8.2.14）式中的 A 矩阵。因此，将 $F(x)$ 的 Jacobi 矩阵记为 $F'(x)$ 是极自然的。当然，反过来Jacobi矩阵的存在并不能保证 $F(x)$ 为 F-可微。例如，当 $m=1$ 时，即 $F(x)$ 为 n 元函数时，假定在点 x 它的所有偏导数均存在，那末它的 Jacobi 矩阵即为

$$F'(x)=\left(\frac{\partial F}{\partial \xi_1},\ \frac{\partial F}{\partial \xi_2},\cdots,\ \frac{\partial F}{\partial \xi_n}\right), \tag{8.2.16}$$

但 $F(x)$ 在 x 点未必可微，也就是说，$F(x)$ 在点 x 处未必 F-可微。

定义8.2.2 设 F：$D\subset R^n\to R^m$，如果 F 在 D 的内点为 F-可微，则把使（8.2.14）式成立的 $A\in R^{m\times n}$ 记为 $F'(x)$，即

$$F'(x)=\begin{pmatrix}\frac{\partial f_1}{\partial \xi_1} & \frac{\partial f_1}{\partial \xi_2} & \cdots & \frac{\partial f_1}{\partial \xi_n}\\ \frac{\partial f_2}{\partial \xi_1} & \frac{\partial f_2}{\partial \xi_2} & \cdots & \frac{\partial f_2}{\partial \xi_n}\\ \cdots & \cdots & \cdots & \\ \frac{\partial f_m}{\partial \xi_1} & \frac{\partial f_m}{\partial \xi_2} & \cdots & \frac{\partial f_m}{\partial \xi_n}\end{pmatrix}, \tag{8.2.17}$$

并称它为 $F(x)$ 在 x 处的 F-导数。

至此我们即可明了：函数 $F(x)$ 满足假设1（a）实际上就是指 $F(x)$ 在 D 内 F-可微。事实上，如果函数 $f_i(\xi_1,\cdots,\xi_n)$（$i=1,\cdots,n$）在 D 内连续且有连续的一阶偏导数，则对一切 i 成立

$$\lim_{h\to 0}\frac{f_i(x+h)-f_i(x)-\nabla f_i(x)^T h}{\|h\|}=0, \tag{8.2.18}$$

因此

$$\lim_{h\to 0}\frac{F(x+h)-F(x)-F'(x)h}{\|h\|}=0, \qquad (8.2.19)$$

而（8.2.19）式等价于

$$\lim_{h\to 0}\left(\frac{1}{\|h\|}\right)\|F(x+h)-F(x)-F'(x)h\|=0.$$

此即 $F(x)$ 在点 x 处 F-可微。

例8.2.1 用牛顿法求非线性方程组

$$\begin{cases} f_1(\xi_1,\ \xi_2)\equiv 4\xi_1^2+\xi_2^2+2\xi_1\xi_2-\xi_2-2=0, \\ f_2(\xi_1,\ \xi_2)\equiv 2\xi_1^2+3\xi_1\xi_2+\xi_2^2-3=0, \end{cases}$$

的解，初始近似取 $x_0=(0.4,\ 0.9)^T$。

据算法8.2.1首先计算 $F(x_0)$：

$$f_1(\xi_1^{(0)},\ \xi_2^{(0)})=-0.73,$$
$$f_2(\xi_1^{(0)},\ \xi_2^{(0)})=-0.79,$$

然后计算 $F'(x_0)$，为此求出

$$f'_{1\xi_1}=8\xi_1+2\xi_2,\ f'_{1\xi_2}=2\xi_1+2\xi_2-1,$$
$$f'_{2\xi_1}=4\xi_1+3\xi_2,\ f'_{2\xi_2}=3\xi_1+2\xi_2.$$

于是

$$F'(x_0)=\begin{pmatrix} 5.0 & 1.6 \\ 4.3 & 3.0 \end{pmatrix}.$$

因此线性方程组 $F'(x_0)\Delta x_0=-F(x_0)$ 为

$$5.0\Delta\xi_1^{(0)}+1.6\Delta\xi_2^{(0)}=0.73,$$
$$4.3\Delta\xi_1^{(0)}+3.0\Delta\xi_2^{(0)}=0.79,$$

解得 $\Delta x=(0.114,\quad 0.100)^T$，故 $x_1=x_0+\Delta x_0=(0.514,\ 1.000)^T$。重复以上计算，得 $F(x_1)=(0.084784,\ 0.070392)^T$，而

$$F'(x_1)=\begin{pmatrix} 6.112 & 2.028 \\ 5.056 & 3.542 \end{pmatrix}.$$

再解线性方程组

$$6.112\Delta\xi_1^{(1)}+2.028\Delta\xi_2^{(1)}=-0.084784,$$
$$5.056\Delta\xi_1^{(1)}+3.542\Delta\xi_2^{(1)}=-0.070392,$$

得 $\Delta\boldsymbol{x}_1=(-0.013826,\quad -0.000138)^T$，$\boldsymbol{x}_2=\boldsymbol{x}_1+\Delta\boldsymbol{x}_1=(0.500174,0.999862)^T$。进而 $\boldsymbol{F}(\boldsymbol{x}_2)=(0.000768,0.000387)^T$。在点 $\boldsymbol{x}_2$ 处的 Jacobi 矩阵为

$$\begin{pmatrix} 6.00116 & 2.000072 \\ 5.000282 & 3.500246 \end{pmatrix}.$$

解线性方程组

$$\begin{cases} 6.00116\Delta\xi_1^{(2)}+2.000072\Delta\xi_2^{(2)}=-0.000768, \\ 5.000282\Delta\xi_1^{(2)}+3.500246\Delta\xi_2^{(2)}=-0.000387, \end{cases}$$

得 $\Delta\boldsymbol{x}_2=(-0.000174,\quad 0.000138)^T$，$\boldsymbol{x}_3=\boldsymbol{x}_2+\Delta\boldsymbol{x}_2=(0.500000,\ 1.000000)^T$。最后，我们发现 $\boldsymbol{F}(\boldsymbol{x}_3)=0$。也就是说经过三次迭代，竟然得到了原方程组的精确解。当然，这是一个特例，一般情形只能得到近似解。但从这个例子可以看出，牛顿法的收敛速度是比较快的。在§4，我们还将具体分析牛顿法的收敛性。

§3 迭代过程的收敛率

本节的主要任务是对迭代序列 $\{\boldsymbol{x}_k\}\subset \boldsymbol{R}^n$ 收敛速度的度量方法予以定义。

定义8.3.1 设 $\{\boldsymbol{x}_k\}\subset \boldsymbol{R}^n$ 是由某迭代算法所生成，它收敛于（8.1.1）的解 $\boldsymbol{x}^*$。若对某向量范数 $\|\cdot\|$，存在一与 k 无关的 $q\in(0,\ 1)$ 以及正整数 k_0，当 $k\geqslant k_0$ 时有

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|\leqslant q\|\boldsymbol{x}_k-\boldsymbol{x}^*\| \qquad (8.3.1)$$

成立，则称迭代序列$\{\boldsymbol{x}_k\}$至少是线性收敛的。

以上定义意味着，当 $k \geqslant k_0$ 时，每经一次迭代，误差至少以因子 q 缩小。一个实用的迭代算法至少应具有线性收敛速度。

一个新的、有价值的局部加速方法，它应该具有下述超线性收敛性。以后我们介绍的大多数算法，在一定条件下均具有超线性收敛的性质。

定义8.3.2 设$\{\boldsymbol{x}_k\} \subset \boldsymbol{R}^n$ 由某迭代算法所生成，它收敛于(8.1.1)的解 $\boldsymbol{x}^*$。若对某种向量范数 $\|\cdot\|$，存在一收敛于零的实数列$\{q_k\}$，使对任何 k 成立

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \leqslant q_k \|\boldsymbol{x}_k - \boldsymbol{x}^*\|, \tag{8.3.2}$$

则称$\{\boldsymbol{x}_k\}$**超线性收敛**，或者更确切地说是商超线性收敛，有时也说该算法是局部超线性收敛。

根据以上定义即可推知：只要当 $k \geqslant k_0$ 且 $\boldsymbol{x}_k \neq \boldsymbol{x}^*$ 时，则$\{\boldsymbol{x}_k\}$超线性收敛就意味着

$$\lim_{k \to \infty} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| / \|\boldsymbol{x}_k - \boldsymbol{x}^*\| = 0. \tag{8.3.3}$$

以下定理是序列$\{\boldsymbol{x}_k\}$超线性收敛的一个必要条件，它对构造某些迭代算法的收敛准则是有用的。

定理8.3.1 设序列$\{\boldsymbol{x}_k\}$超线性收敛于 $\boldsymbol{x}^*$，当 $k \geqslant k_0$ 且 $\boldsymbol{x}_k \neq \boldsymbol{x}^*$ 时，则

$$\lim_{k \to \infty} \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} = 1. \tag{8.3.4}$$

但一般而言，反之未必成立。

证 因为对任何 $k \geqslant k_0$ 均有

$$\left| \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} - \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|} \right| \leqslant \frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|},$$

即

$$\left|\frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|}-1\right| \leqslant \frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|}.$$

据（8.3.3）即可推知（8.3.4）成立。

上述定理的逆命题未必成立，我们给出一个反例。

例8.3.1 设序列$\{\boldsymbol{x}_k\}$定义为

$$\begin{cases} x_{2i-1}=\dfrac{1}{i!}, \\ x_{2i}=2x_{2i-1}, \end{cases} \qquad (i=1, 2, \cdots)$$

又设$\boldsymbol{x}^*=0$，则上述序列收敛于$\boldsymbol{x}^*$，但据序列定义，显然它并不超线性收敛于$\boldsymbol{x}^*$。然而

$$\frac{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|}=\begin{cases} 1 & (k=2i-1, \ i\geqslant 1), \\ 1-\dfrac{1}{2(i+1)} & (k=2i, \ i\geqslant 1). \end{cases}$$

于是，对该序列而言，（8.3.4）是成立的。

定理8.3.1告诉我们，如果已知$\{\boldsymbol{x}_k\}$超线性收敛于$\boldsymbol{x}^*$，那末就可利用$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|$来估计误差$\|\boldsymbol{x}_k-\boldsymbol{x}^*\|$。据此，即可构成如下迭代收敛准则：事先指定正小数$\varepsilon_1$，如果有

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|\leqslant\varepsilon_1\|\boldsymbol{x}_k\| \tag{8.3.5}$$

或

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|\leqslant\varepsilon_1\|\boldsymbol{x}_{k+1}\| \tag{8.3.6}$$

成立，那末迭代终止。有时候以上迭代收敛准则常和

$$\|\boldsymbol{F}(\boldsymbol{x}_k)\|<\varepsilon_2 \tag{8.3.7}$$

联用，以防止出现只顾一头的偏向，这里ε_2也为事先指定的正小数（当然还可以采取其它措施）。这样做的目的是，在实际计算时，$\{\boldsymbol{x}_k\}$并不一定超线性收敛。其次，即便$\{\boldsymbol{x}_k\}$超线性收敛，$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|$也未必为$k$的单调减函数。

§4 牛顿法的收敛性分析

牛顿法的主要优点可见于以下二个定理。

定理8.4.1 （牛顿法的局部收敛性）

设 $F: R^n \to R^n$ 满足假设 1，则存在一闭球 $\overline{S(x^*, \delta)} \subset D$，对任何 $x_0 \in \overline{S(x^*, \delta)}$，牛顿迭代解

$$x_{k+1} = x_k - F'(x_k)^{-1} F(x_k)$$

对任何 k 均有定义且恒不越出该球。序列 $\{x_k\}$ 超线性收敛于 x^*。

定理8.4.2 （牛顿法的平方收敛性）

设 $F: R^n \to R^n$ 满足假设 1，$F'(x)$ 在点 x^* 满足 Lipschitz 条件，亦即存在正常数 κ 使

$$\|F'(x) - F'(x^*)\| \leqslant \kappa \|x - x^*\|, \quad \forall x \in D \qquad (8.4.1)$$

成立，则存在正常数 q 使

$$\|x_{k+1} - x^*\| \leqslant q \|x_k - x^*\|^2 \quad (k = 0, 1, \cdots) \qquad (8.4.2)$$

成立，即 $\{x_k\}$ 至少是平方收敛的。

请注意：和一维情形类似，可以证明，如果 D 充分小，而 $F(x)$ 在 x^* 处二次可微*，那末（8.4.1）即能满足。所谓 $F(x)$ 在 x 处二次可微，这里是指 $F(x)$ 的所有分量函数 $f_i(x)$ $(i = 1, \cdots, n)$ 在 x 处二次连续可微。

---

*）仿照定义8.2.1，对二阶导数可作如下定义：

设 $F: D \subset R^n \to R^n$ 在 D 内 Fréchet 可微，x 为 D 的内点，若存在双线性算子 $A: R^n \times R^n \to R^n$ 使

$$\lim_{h \to 0} \left(\frac{1}{\|h\|}\right) \|F'(x+h) - F'(x) - Ah\| = 0$$

或

$$F'(x+h) = F'(x) + Ah + o(\|h\|)$$

对所有 $h \in R^n$ 成立，则 $F(x)$ 在点 x 处二次可微并称算子 A 为 F 在点 x 处的二阶导数，记 $A = F''(x)$。

今后我们称 $\boldsymbol{F}(\boldsymbol{x})$ 在 D 上满足（8.4.1）为假设2。

为了证明以上二个定理，我们首先证明下面几个引理。

引理8.4.1 设矩阵 $\boldsymbol{A}(\boldsymbol{x})\in \boldsymbol{R}^{n\times n}$ 在点 $\boldsymbol{x}^*\in D$ 处连续且 $\boldsymbol{A}(\boldsymbol{x}^*)$ 为非异，则存在闭球 $\overline{S(\boldsymbol{x}^*, \delta)}\subset D$ 及正数 γ，使对一切 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*, \delta)}$，$\boldsymbol{A}(\boldsymbol{x})$ 亦为非异且

$$\|\boldsymbol{A}(\boldsymbol{x})^{-1}\|\leqslant\gamma.$$

证 令 $\alpha=\|\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\|$，取 $\beta<\dfrac{1}{\alpha}$，对取定的 β 选取这样的 $\delta>0$，以保证 $\overline{S(\boldsymbol{x}^*, \delta)}\subset D$ 且当 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*, \delta)}$ 时有

$$\|\boldsymbol{A}(\boldsymbol{x}^*)-\boldsymbol{A}(\boldsymbol{x})\|\leqslant\beta$$

成立。据 $\boldsymbol{A}(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处的连续性，这是能办到的。考虑到当 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*, \delta)}$ 时有

$$\|\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})\|\leqslant\|\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\|\,\|\boldsymbol{A}(\boldsymbol{x}^*)-\boldsymbol{A}(\boldsymbol{x})\|\leqslant\alpha\beta<1, \tag{8.4.3}$$

又

$$\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})=\boldsymbol{I}-(\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})),$$

因此，当 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*, \delta)}$ 时，$\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})$ 必为非异，于是 $\boldsymbol{A}(\boldsymbol{x})$ 亦必为非异。又当 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*, \delta)}$ 时应有

$$\begin{aligned}\|\boldsymbol{A}(\boldsymbol{x})^{-1}\|&=\|[\boldsymbol{I}-(\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})]^{-1}\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\|\\&\leqslant\|\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\|\,\|[\boldsymbol{I}-(\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})]^{-1}\|\\&\leqslant\|\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\|\frac{1}{1-\|\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{x}^*)^{-1}\boldsymbol{A}(\boldsymbol{x})\|}\\&\leqslant\frac{\alpha}{1-\alpha\beta}=\gamma.\end{aligned}$$

有时以上引理常以下面形式出现：设 $\boldsymbol{A}_k\in\boldsymbol{R}^{n\times n}$ 而 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$ 非异且

$$\|\boldsymbol{A}_k-\boldsymbol{A}\|\leqslant\beta,$$

若$\|A^{-1}\|\leqslant\alpha$且$\alpha\beta<1$，则$A_k$亦为非异且

$$\|A_k^{-1}\|\leqslant\frac{\alpha}{1-\alpha\beta}. \tag{8.4.4}$$

文献上常称这个引理为 **Banach** 引理，证明过程则完全和引理8.4.1相同。**Banach** 引理的意义就是：若A_k能接近一个可逆阵，那末A_k本身也是可逆的。

引理8.4.2 设φ：$[a, b]\subset R^1\to R^n$连续，则有

$$\left\|\int_a^b\varphi(t)dt\right\|\leqslant\int_a^b\|\varphi(t)\|dt. \tag{8.4.5}$$

证：由于$\|\varphi(t)\|$是t的连续函数，故$\|\varphi(t)\|$黎曼可积。据向量函数的积分定义知，对任意给定的$\varepsilon>0$，存在分划$a=t_0<\cdots<t_m=b$使

$$\left\|\int_a^b\varphi(t)dt-\sum_{i=1}^m\varphi(t_i)(t_i-t_{i-1})\right\|\leqslant\varepsilon. \tag{8.4.6}$$

另一方面，据一元函数积分定义有

$$\left|\int_a^b\|\varphi(t)\|dt-\sum_{i=1}^m\|\varphi(t_i)\|(t_i-t_{i-1})\right|\leqslant\varepsilon. \tag{8.4.7}$$

由（8.4.6）和（8.4.7）即可推知

$$\begin{aligned}\left\|\int_a^b\varphi(t)dt\right\|&\leqslant\left\|\sum_{i=1}^m\varphi(t_i)(t_i-t_{i-1})\right\|+\varepsilon\\&\leqslant\sum_{i=1}^m\|\varphi(t_i)\|(t_i-t_{i-1})+\varepsilon\\&\leqslant\int_a^b\|\varphi(t)\|dt+2\varepsilon,\end{aligned}$$

因为ε为任意小的正数，故（8.4.5）成立。

中值定理在一元函数或多元函数微积分中占有十分重要的地位。对于向量函数的微积分情况也是这样。不过，它的中值定理的形式略有变化。

引理8.4.3 设F：$R^n\to R^n$满足假设1(1)，则对任意x，$y\in D$有不等式

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{z})\|\leqslant\max_{0\leqslant t\leqslant 1}\|\boldsymbol{F}'[z+t(\boldsymbol{y}-\boldsymbol{z})]\|\|\boldsymbol{z}-\boldsymbol{y}\| \quad (8.4.8)$$

成立。

证 由于D为开的凸集，故对$0\leqslant t\leqslant 1$，$\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z})\in D$，因此$\boldsymbol{F}'(\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z}))$有定义。对任何$\boldsymbol{y}$，$\boldsymbol{z}\in D$以下等式显然成立。

$$\begin{aligned}f_i(\boldsymbol{y})-f_i(\boldsymbol{z})&=\int_0^1\frac{d}{dt}f_i[\zeta_1+t(\eta_1-\zeta_1),\cdots,\ \zeta_n+t(\eta_n-\zeta_n)]dt\\&=\int_0^1\sum_{j=1}^n\frac{\partial}{\partial\xi_j}f_i[\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z})](\eta_j-\zeta_j)dt\\&=\int_0^1 f_i'[\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z})](\boldsymbol{y}-\boldsymbol{z})dt,\end{aligned}$$

于是得到$\boldsymbol{F}(\boldsymbol{x})$积分形式的中值定理：

$$\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{z})=\int_0^1\boldsymbol{F}'[\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z})](\boldsymbol{y}-\boldsymbol{z})dt, \quad (8.4.9)$$

两端取范数并利用不等式（8.4.5），即可推得（8.4.8）。

引理8.4.4 设$\boldsymbol{F}$：$\boldsymbol{R}^n\to\boldsymbol{R}^n$满足假设1(1)，则对任何$\boldsymbol{x}$，$\boldsymbol{y}$，$\boldsymbol{z}\in D$有以下不等式

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{z})-\boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{z})\|\leqslant\max_{0\leqslant t\leqslant 1}\|\boldsymbol{F}'[\boldsymbol{z}+t(\boldsymbol{y}-\boldsymbol{z})]-\boldsymbol{F}'(\boldsymbol{x})\|\|\boldsymbol{y}-\boldsymbol{z}\| \quad (8.4.10)$$

成立。进而，如果$\boldsymbol{F}'(\boldsymbol{x})$在$D$中Lipschitz连续，即对任何$\boldsymbol{x}$，$\boldsymbol{y}\in D$，存在正常数$\kappa$，使

$$\|\boldsymbol{F}'(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{y})\|\leqslant\kappa\|\boldsymbol{x}-\boldsymbol{y}\| \quad (8.4.11)$$

成立，则

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{z})-\boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{z})\|\leqslant\kappa\sigma(\boldsymbol{y},\boldsymbol{z})\|\boldsymbol{y}-\boldsymbol{z}\|, \quad (8.4.12)$$

其中$\sigma(\boldsymbol{y},\boldsymbol{z})=\max\{\|\boldsymbol{y}-\boldsymbol{x}\|,\ \|\boldsymbol{z}-\boldsymbol{x}\|\}$。特别地，当$\boldsymbol{z}=\boldsymbol{x}$时有不等式

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|\leqslant\frac{1}{2}\kappa\|\boldsymbol{y}-\boldsymbol{x}\|^2 \quad (8.4.13)$$

成立，这里 κ 为 Lipschitz 常数。

证 利用（8.4.9）和引理8.4.2可推得

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|=\|\int_0^1\{\boldsymbol{F}'[\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})]-\boldsymbol{F}'(\boldsymbol{x})\}(\boldsymbol{y}-\boldsymbol{x})dt\|$$

$$\leqslant\int_0^1\|\boldsymbol{F}'[\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})]-\boldsymbol{F}'(\boldsymbol{x})\|\|\boldsymbol{y}-\boldsymbol{x}\|dt$$

$$\leqslant\max_{0\leqslant t\leqslant 1}\|\boldsymbol{F}'[\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})]-\boldsymbol{F}'(\boldsymbol{x})\|\|\boldsymbol{y}-\boldsymbol{x}\|.$$

如果 $\boldsymbol{F}'(\boldsymbol{x})$ 在 D 中 Lipschitz 连续，那末可进一步推得

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|\leqslant\kappa\int_0^1\|[\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})]-\boldsymbol{x}\|\|\boldsymbol{y}-\boldsymbol{x}\|dt$$

$$\leqslant\kappa\max_{0\leqslant t\leqslant 1}\|\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})-\boldsymbol{x}\|\|\boldsymbol{y}-\boldsymbol{x}\|$$

$$=\kappa\sigma(\boldsymbol{y},\boldsymbol{x})\|\boldsymbol{y}-\boldsymbol{x}\|.$$

而

$$\|\boldsymbol{F}(\boldsymbol{y})-\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|\leqslant\kappa\|\boldsymbol{y}-\boldsymbol{x}\|^2\int_0^1 tdt$$

$$=\frac{1}{2}\kappa\|\boldsymbol{y}-\boldsymbol{x}\|^2.$$

至此即能证明早已提出的定理8.4.1和定理8.4.2.

定理8.4.1（牛顿法的局部收敛性）的证明：

据定理8.4.1假定，$\boldsymbol{F}'(\boldsymbol{x})$ 在点 $\boldsymbol{x}^*\in D$ 处连续且 $\boldsymbol{F}'(\boldsymbol{x}^*)$ 为非异，故利用引理 8.4.1 知：存在闭球 $\overline{S(\boldsymbol{x}^*,\ \delta')}\subset D$ 及正数 γ，对一切 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*,\ \delta')}$，$\boldsymbol{F}'(\boldsymbol{x})$ 为非异且 $\|\boldsymbol{F}'(\boldsymbol{x})^{-1}\|\leqslant\gamma$.

另一方面，据 $\boldsymbol{F}'(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处的连续性知，对任给定的 $\varepsilon>0$ 总存在 $\delta>0(\delta\leqslant\delta')$，对一切 $\boldsymbol{x}\in\overline{S(\boldsymbol{x}^*,\ \delta)}$ 有

$$\|\boldsymbol{F}'(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x}^*)\|\leqslant\varepsilon$$

成立。又因 $\boldsymbol{F}(\boldsymbol{x})$ 于 $\boldsymbol{x}^*$ 处 F-可微，我们可将 δ 选取充分小，使 $2\varepsilon\gamma=q<1$ 且不等式

$$\|\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}(\boldsymbol{x}^*)-\boldsymbol{F}'(\boldsymbol{x}^*)(\boldsymbol{x}-\boldsymbol{x}^*)\|\leqslant\varepsilon\|\boldsymbol{x}-\boldsymbol{x}^*\|$$

$$\forall \boldsymbol{x}\in\overline{S(\boldsymbol{x}^*,\ \delta)}$$

成立。

现设 $\boldsymbol{x}\in S(\boldsymbol{x}^*,\ \delta)$ 并记 $G(\boldsymbol{x})=\boldsymbol{x}-\boldsymbol{F}'(\boldsymbol{x})^{-1}\boldsymbol{F}(\boldsymbol{x})$，则有

$$\begin{aligned}\|G(\boldsymbol{x})-\boldsymbol{x}^*\|&=\|\boldsymbol{x}-\boldsymbol{x}^*-\boldsymbol{F}'(\boldsymbol{x})^{-1}[\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}(\boldsymbol{x}^*)]\|\\&\leqslant\|-\boldsymbol{F}'(\boldsymbol{x})^{-1}\|[\|\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}(\boldsymbol{x}^*)\\&\quad-\boldsymbol{F}'(\boldsymbol{x}^*)(\boldsymbol{x}-\boldsymbol{x}^*)\|\\&\quad+\|\boldsymbol{F}'(\boldsymbol{x})-\boldsymbol{F}'(\boldsymbol{x}^*)\|\|\boldsymbol{x}-\boldsymbol{x}^*\|]\\&\leqslant2\varepsilon\gamma\|\boldsymbol{x}-\boldsymbol{x}^*\|=q\|\boldsymbol{x}-\boldsymbol{x}^*\|.\end{aligned}\tag{8.4.14}$$

现若取 $\boldsymbol{x}_0$ 使 $\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\leqslant\delta$，则据(8.4.14)且记 $\boldsymbol{x}_1=G(\boldsymbol{x}_0)$，有

$$\|\boldsymbol{x}_1-\boldsymbol{x}^*\|\leqslant q\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\leqslant q\delta<\delta.$$

由此可见 $\boldsymbol{x}_1\in\overline{S(\boldsymbol{x}^*,\ \delta)}$。利用归纳法即可证

$$\|\boldsymbol{x}_k-\boldsymbol{x}^*\|\leqslant q\|\boldsymbol{x}_{k-1}-\boldsymbol{x}^*\|<\delta\tag{8.4.15}$$

对所有的 $k\geqslant1$ 均成立，这里 $\boldsymbol{x}_k=G(\boldsymbol{x}_{k-1})$。逐次利用这些不等式即可得到

$$\|\boldsymbol{x}_k-\boldsymbol{x}^*\|\leqslant q^k\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\qquad(k=1,\ 2,\cdots),\tag{8.4.16}$$

又因 $0<q<1$，故

$$\lim_{k\to\infty}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|=0,$$

亦即

$$\lim_{k\to\infty}\boldsymbol{x}_k=\boldsymbol{x}^*.$$

这样，我们就证明了牛顿法的局部收敛性。

再令

$$\gamma\left[\frac{\|\boldsymbol{F}(\boldsymbol{x}_k)-\boldsymbol{F}(\boldsymbol{x}^*)-\boldsymbol{F}'(\boldsymbol{x}^*)(\boldsymbol{x}_k-\boldsymbol{x}^*)\|}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|}\right.$$
$$\left.+\|\boldsymbol{F}'(\boldsymbol{x}_k)-\boldsymbol{F}'(\boldsymbol{x}^*)\|\right]=q_k,$$

则据(8.4.14)即有

$$\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|\leqslant q_k\|\boldsymbol{x}_k-\boldsymbol{x}^*\|.\qquad(k=0,\ 1,\cdots)$$

据 $F(x)$ 在 x^* 的 F-可微性以及 $F'(x)$ 在 x^* 处的连续性即可推知

$$\lim_{k\to\infty} q_k = 0,$$

故牛顿迭代序列 $\{x_k\}$ 是局部超线性收敛的。

定理8.4.2的证明：

由不等式

$$\|x_{k+1} - x^*\| \leqslant \|F'(x_k)^{-1}\| \|F(x_k) - F(x^*) - F'(x_k)(x_k - x^*)\|,$$

并用引理8.4.4即可推得

$$\|x_{k+1} - x^*\| \leqslant \frac{3}{2}\gamma\kappa\|x_k - x^*\|^2 \qquad (k = 0, 1, \cdots),$$

由此可见 $\{x_k\}$ 至少是平方收敛。

至此，我们有必要对定理 8.4.1 的证明予以回顾。牛顿迭代法实际上可看作形如

$$x = G(x) \tag{8.4.17}$$

的非线性方程组的简单迭代法，即

$$x_{k+1} = G(x_k) \qquad (k = 0, 1, \cdots). \tag{8.4.18}$$

定理8.4.1主要论证了函数 $G(x)$ 满足以下三个条件：

(1) 存在一闭球 $S(x^*, \delta)$，函数 $G(x)$ 在该闭球上有定义；

(2) $G(\overline{S(x^*, \delta)}) \subset S(x^*, \delta)$；

(3) 存在 $q \in (0, 1)$ 使

$$\|G(x) - G(x^*)\| \leqslant q\|x - x^*\|, \ \forall x \in \overline{S(x^*, \delta)}. \tag{8.4.19}$$

第三个条件通常称为**压缩条件**。在定理 8.4.1 中这个条件实际上是

$$G'(x^*) = 0 \tag{8.4.20}$$

或

$$\|G'(\boldsymbol{x}^*)\| = 0,$$

这可从（8.4.14）看出。一般地，若 $G(\boldsymbol{x})$ 在 $\boldsymbol{x}^*$ 处连续可微且

$$\rho(G'(\boldsymbol{x}^*)) < 1, \qquad (8.4.21)$$

即能保证存在球 $\overline{S(\boldsymbol{x}^*, \delta)}$，$G(\boldsymbol{x})$ 在 $\overline{S(\boldsymbol{x}^*, \delta)}$ 上满足压缩条件。事实上，据范数理论知：对任意给定的 $\varepsilon>0$，存在某种范数 $\|\cdot\|_k$ 使

$$\|G'(\boldsymbol{x}^*)\|_k \leqslant \rho(G'(\boldsymbol{x}^*)) + \varepsilon, \qquad (8.4.22)$$

而对 $\varepsilon>0$ 又存在 $\overline{S(\boldsymbol{x}^*, \delta)}$，当 $\boldsymbol{x} \in \overline{S(\boldsymbol{x}^*, \delta)}$ 时有

$$\begin{aligned}\|\boldsymbol{\varphi}(\boldsymbol{x}) - G(\boldsymbol{x}^*)\|_k &\leqslant \int_0^1 \|G'[\boldsymbol{x}^* + t(\boldsymbol{x} - \boldsymbol{x}^*)]\|_k \|\boldsymbol{x} - \boldsymbol{x}^*\|_k dt \\ &\leqslant (\|G'(\boldsymbol{x}^*)\|_k + \varepsilon)\|\boldsymbol{x} - \boldsymbol{x}^*\|_k \qquad (8.4.23)\end{aligned}$$

成立，从而

$$\|G(\boldsymbol{x}) - G(\boldsymbol{x}^*)\|_k \leqslant (\rho(\boldsymbol{\varphi}'(\boldsymbol{x}^*)) + 2\varepsilon)\|\boldsymbol{x} - \boldsymbol{x}^*\|_k.$$

由于 ε 的任意性，故总有 $q = (\rho(\boldsymbol{\varphi}'(\boldsymbol{x}^*)) + 2\varepsilon) < 1$ 成立，即 $G(\boldsymbol{x})$ 在 $\overline{S(\boldsymbol{x}^*, \delta)}$ 上满足压缩条件。

从以上几个定理可以看出：牛顿法有二大优点（Brown 方法亦类似）。其一，在一般条件下牛顿法是超线性收敛的；如果再加上假设 2 这个条件，那末牛顿法至少是平方收敛的。粗略地说，后者意味着牛顿法每迭代一次，有效数位基本上增加了一倍，这就是说，如果由准确到一位的近似解出发，那只要经四次迭代就可得到准确到十六位的近似解。其二，牛顿法在一般条件下总存在一个吸引域 S。换言之，只要初始近似解落在 S 中，那么由此出发，生成的迭代序列 $\{\boldsymbol{x}_k\}$ 总不会越出域 S，而且收敛于 $\boldsymbol{x}^*$。

另外也应指出：牛顿法是自动校正的，这就是说 $\boldsymbol{x}_{k+1}$ 仅仅依赖于 $\boldsymbol{F}$ 和 $\boldsymbol{x}_k$，因而前面迭代时产生的舍入误差不会一步

步传下去。牛顿法这一优点我们将会看到并不能为拟牛顿法所共有。

正如我们在引言中早就指出的，确定一个初始近似解$\boldsymbol{x}_0$有时是很困难的，这正是牛顿法的一大缺点（Brown 方法也一样）。因为对某一具体问题，吸引区域 S 很可能极小，因此初始近似解必须选得很好，然而这常常是很难办到的。克服这一缺点的措施之一，可以在牛顿公式中引入收敛因子 $\lambda_k>0$，以扩大收敛区域。这时迭代公式具有如下形式：

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\lambda_k\boldsymbol{F}'(\boldsymbol{x}_k)^{-1}\boldsymbol{F}(\boldsymbol{x}_k)\qquad(k=0,\ 1,\ \cdots)\tag{8.4.24}$$

而 λ_k 如此选择，以保证 $\{\|\boldsymbol{F}(\boldsymbol{x}_k)\|\}$为严格单调下降序列。选取的办法通常使 $\lambda_k>0$ 为极值问题

$$\min_{\lambda>0}\|\boldsymbol{F}[\boldsymbol{x}_k-\lambda\boldsymbol{F}'(\boldsymbol{x}_k)^{-1}\boldsymbol{F}(\boldsymbol{x}_k)]\|\tag{8.4.25}$$

的解。设 $\varphi(\lambda)\equiv\|\boldsymbol{F}[\boldsymbol{x}_k-\lambda\boldsymbol{F}'(\boldsymbol{x}_k)^{-1}\boldsymbol{F}(\boldsymbol{x}_k)]\|$ 为区间$[a,\ b]$上的严格单峰函数*)，则极小点 λ_k 可采用0.618优选法寻找。0.618搜索方法的步骤如下：

在 λ_k 的预估区间 $[a,\ b]$ 中，计算两个试验点 $a_1=a+0.381966(b-a)$，$b_1=a+0.618034(b-a)$ 以及其上的函数值 φ_1 和 φ_2。如果 $\varphi_1<\varphi_2$，则取$[a,\ b_1]$为新的估值区间，否则取$[a_1,\ b]$。然后重复上述过程。由于试验点取法特殊，下一次要计算的两点之一可由前次算出试验点近似替代，而另外一点又近似等于区间端点之和减去已知的试验点。所以，除第一次需要计算两个点上的函数值外，以后只要计算一个新点上的函数

---

*) 一个函数 $\varphi:[a,\ b]\subset R^1\to R^1$ 在$[a,\ b]$上是严格单峰的，如果存在$\lambda^*\in[a,\ b]$为 $\varphi(\lambda)$ 在$[a,\ b]$上的整体极小点，并且对任何 λ'，$\lambda''\in[a,\ b]$，当 $\lambda'<\lambda''$ 时，若 $\lambda''\leqslant\lambda^*$，则 $\varphi(\lambda')>\varphi(\lambda'')$；若 $\lambda^*\leqslant\lambda'$，则 $\varphi(\lambda'')>\varphi(\lambda')$。

值。当区间长度比原始区间长度小到0.01倍时，搜索结束并取其中点作为 λ_k。λ_0 的预估区间可取[0, $1.618034\lambda_{-1}$]，其中λ_{-1}事先选定，可取0.1或1。λ_k 的预估区间可取[0, $1.618034\lambda_{k-1}$]，这里 λ_{k-1} 为前次迭代的收敛因子。

牛顿法另一个缺点（Brown 方法也类似）是：对每个 k 都必须计算矩阵 $F'(x_k)$，向量 $F(x_k)$ 以及 $O(n^3)$ 次算术运算求解线性方程组。这就是说，每一步都要计算 n^2 个偏导函数值。对于大多数 $F(x)$ 而言，完成它将花费很大的工作量。因此，如果对某一问题的 Jacobi 矩阵相对而言计算量不太大，如Jacobi矩阵为大型稀疏矩阵，那末牛顿法是值得采用的；如果对某一问题的 Jacobi 矩阵的计算量很大，那末可选用各种修正的牛顿法或其它方法。例如，为了减少牛顿法中矩阵求逆的次数，一个较为普遍采用的技巧是：在一定的迭代次数内，使 Jacobi矩阵固定，这对于 Jacobi 矩阵变化很慢的情形是十分有效的。具体说，定义 $x_{k+1}=y_{k,k_0}$，$y_{k,0}=x_k$ 而

$$y_{k,j}=y_{k,j-1}-F'(x_k)^{-1}F(x_{k,j-1})\quad(j=1,\ 2,\cdots,\ k_0). \tag{8.4.26}$$

可以证明由(8.4.26)产生的迭代序列 $\{x_k\}$收敛阶*至少为 k_0+1（习题8.7），所以它也是产生高阶迭代过程的一个有效且简单的途径。

最后还必须指出：如果 $F'(x^*)$ 奇异，那末牛顿法的收敛性质就要被破坏。近年来，对于奇异点的牛顿法已有许多研究成果。

---

*) 设序列 $\{x_k\}$ 收敛于 x^*，则$\{x_k\}$的收敛阶为

$$p=\sup\{q\in R^1\mid \overline{\lim_{k\to\infty}}\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|^q}\in[0,\ \infty)\text{而}q\geqslant1\}.$$

§5 离散的牛顿法

由前节分析我们知道，牛顿法的主要缺点是对每个 k 都必须计算矩阵 $F'(x_k)$，亦即对每个 k 都必须计算 n^2 个导数。因此，人们很自然地会想到：矩阵 $F'(x_k)$ 中 (i, j) 元用一个差商来近似替代。换言之，将 Jacobi 矩阵离散化。通常最简单的办法是，$\frac{\partial f_i}{\partial \xi_j}$ 以一阶差商

$$\frac{f_i(x_k + h_k e_j) - f_i(x_k)}{h_k}$$

替代之。其中 e_j 表示第 j 个单位向量，h_k 常取作 $O(\|F(x_k)\|)$ 或 $O(|f_1(x_k)|)$。在此基础上可构成一类离散型牛顿算法 N_k。

算法8.5.1（N_k 算法）

1）对 $k=0, 1, \cdots, m$。

1.1）对充分小的步长 h_k，计算矩阵 J_k，它的第 j 列为

$$J_k e_j = \frac{F(x_k + h_k e_j) - F(x_k)}{h_k} \quad (j=1, \cdots, n), \qquad (8.5.1)$$

1.2）用近似 Jacobi 矩阵 J_k（假定它为非异）执行 k_0 次“牛顿迭代”。亦即定义 $x_{k+1} = y_{k,k_0}$，$y_{k,0} = x_k$ 而

$$y_{k,j} = y_{k,j-1} - J_k^{-1} F(y_{k,j-1}) \ (j=1, \cdots, k_0), \quad (8.5.2)$$

1.3）如果 x_{k+1} 满足精度要求，则转3)。

1.4）计算 $F(x_{k+1})$ 以及 $F(x_{k+1} + h_{k+1} e_j)$（$j=1, \cdots, n$）。

2）NEXT k。

3）置 $x^* = x_{k+1}$。

4）END.

与习题8.7类似，在一定条件下，可以证明 N_k 算法的收敛阶至少为 k_0+1。顺便指出，N_k 算法每迭代一次需计算 $n+k_0$ 次函数值，（这里是指向量函数 $\boldsymbol{F}(\boldsymbol{x})$ 的函数值，其每次计算应包括 n 次纯量函数值）。

很自然 Brown 算法同样也能离散化。离散 Brown 算法的主要步骤如下：

第一步：设 $\boldsymbol{x}_k$ 为（8.1.1）解 $\boldsymbol{x}^*$ 的近似值，则对函数 f_1 在 $\boldsymbol{x}_k$ 的近傍用线性函数

$$f_1(\boldsymbol{x}_k)+f_{1\xi_1;hx}(\boldsymbol{x}_k)(\xi_1-\xi_1^{(k)})+\cdots+f_{1\xi_n;hx}(\boldsymbol{x}_k)(\xi_n-\xi_n^{(k)}) \tag{8.5.3}$$

近似替代。其中 $\boldsymbol{x}=(\xi_1,\cdots,\xi_n)^T$，$\boldsymbol{x}_k=(\xi_1^{(k)},\cdots,\xi_n^{(k)})^T$而

$$f_{1\xi_j;hx}(\boldsymbol{x}_k)=\frac{f_1(\boldsymbol{x}_k+h_j^{(k)}\boldsymbol{e}_j)-f_1(\boldsymbol{x}_k)}{h_j^{(k)}}. \tag{8.5.4}$$

（以下为书写简便计，略去了 h 的上下标，请读者注意）。如果 $\boldsymbol{x}_k$ 充分接近 $\boldsymbol{x}^*$，那末可令(8.5.3)式为零，并由此解出一个变量。譬如说，如果 $f_{1\xi_n;h}(\boldsymbol{x}_k)$ 按模最大，则就可解出 ξ_n 来，即

$$\xi_n=\xi_n^{(k)}-\sum_{j=1}^{n-1}(f_{1\xi_j;h}^{(k)}/f_{1\xi_n;h}^{(k)})(\xi_j-\xi_j^{(k)})-f_1^{(k)}/f_{1\xi_n;h}^{(k)}, \tag{8.5.5}$$

其中 $f_{1\xi_j;h}^{(k)}\equiv f_{1\xi_j;h}(\boldsymbol{x}_k)$，$f_1^{(k)}\equiv f_1(\boldsymbol{x}_k)$。可以证明，$\boldsymbol{F}(\boldsymbol{x})$ 在假设1的条件下，至少有一个非零偏导数 $\frac{\partial f_1(\boldsymbol{x}_k)}{\partial \xi_j}$。因此，相应的近似偏导数 $f_{1\xi_j;h}^{(k)}$ 亦为非零。这样（8.5.5）总是有意义的。选取绝对值为最大的近似偏导数相除，其作用类似于用高斯消去法求解线性方程组时，采用部分主元素方案，以保证方法的数值稳定性。由（8.5.5）式可以看出，ξ_n 乃是 $n-1$ 个变量 ξ_1，$\xi_2,\cdots,\xi_{n-1}$ 的线性函数。为书写清楚起见，记（8.5.5）的左端为 $L_n(\xi_1,\cdots,\xi_{n-1})$且定义 $L_n^{(k)}\equiv L_n(\xi_1^{(k)},\cdots,\xi_{n-1}^{(k)})$。

第二步：对函数 f_2，定义新函数

$$g_2(\xi_1,\cdots,\xi_{n-1})\equiv f_2(\xi_1,\cdots,\xi_{n-1},\ L_n(\xi_1,\cdots,\xi_{n-1}))\quad(8.5.6)$$

且记 $g_2^{(k)}\equiv f_2(\xi_1^{(k)},\cdots,\xi_{n-1}^{(k)},L_n^{(k)})$。注意这时 g_2 是变量 $\xi_1,\cdots,\xi_{n-1}$ 的函数. 同样地，在点 $(\xi_1^{(k)},\cdots,\xi_{n-1}^{(k)})$ 近傍以线性函数近似替代 g_2，并由此解出一个变量. 设 ξ_{n-1} 对应的近似偏导数 $g_{2\xi_{n-1};h}$ 按绝对值为最大，于是解得 ξ_{n-1} 为：

$$\xi_{n-1}=\xi_{n-1}^{(k)}-\sum_{j=1}^{n-2}(g_{2\xi_j;h}^{(k)}/g_{2\xi_{n-1};h}^{(k)})(\xi_j-\xi_j^{(k)})-g_2^{(k)}/g_{2\xi_{n-1};h}^{(k)},\quad(8.5.7)$$

这里近似偏导数 $g_{2\xi_j;h}^{(k)}$ 为

$$g_{2\xi_j;h}^{(k)}\equiv\frac{g_2(\xi_1^{(k)},\cdots,\xi_{j-1}^{(k)},\xi_j^{(k)}+h,\xi_{j+1}^{(k)},\cdots,\xi_{n-1}^{(k)})-g_2^{(k)}}{h}.\quad(8.5.8)$$

由（8.5.7）知，ξ_{n-1} 为其余 $n-2$ 个变量的线性函数。现将此线性函数记作 $L_{n-1}(\xi_1,\cdots,\xi_{n-2})$，同样定义 $L_{n-1}^{(k)}\equiv L_{n-1}(\xi_1^{(k)},\cdots,\xi_{n-2}^{(k)})$。请注意：此时应把 L_n 中的 ξ_{n-1} 换为 L_{n-1}，即此时 L_n 的变量为 $\xi_1,\cdots,\xi_{n-2},L_{n-1}(\xi_1,\cdots,\xi_{n-2})$。

继续进行这种“逐步代入”的过程，至第 $i+1$ 步，得到线性表达式

$$L_{n-i}\equiv\xi_{n-i}^{(k)}-\sum_{j=1}^{n-i-1}(g_{i+1,\xi_j;h}^{(k)}/g_{i+1,\xi_{n-i};h}^{(k)})(\xi_j-\xi_j^{(k)})-g_{i+1}^{(k)}/g_{i+1,\xi_{n-i};h}^{(k)},\quad(8.5.9)$$

其中

$$g_{i+1}^{(k)}\equiv f_{i+1}(\xi_1^{(k)},\cdots,\xi_{n-i}^{(k)},L_{n-i+1}^{(k)},\cdots,L_n^{(k)}),\quad(8.5.10)$$

$$g_{i+1,\xi_j;h}^{(k)}\equiv\frac{g_{i+1}(\xi_1^{(k)},\cdots,\xi_{j-1}^{(k)},\xi_j^{(k)}+h,\xi_{j+1}^{(k)},\cdots,\xi_{n-i}^{(k)})-g_{i+1}^{(k)}}{h}$$

$$(j=1,\cdots,n-i).\quad(8.5.11)$$

由此可见，在完成第 $i+1$ 步时，一个主要的工作就是要计算 g_{i+1} 在点 $(\xi_1^{(k)},\cdots,\xi_{n-i}^{(k)})^T\equiv x_{n-i}^{(k)}$ 以及点 $x_{n-i}^{(k)}+he_j\ (j=1,\cdots,n-i)$

处的值。请注意：虽然计算 g_{i+1} 就是计算 f_{i+1}，但这时 f_{i+1} 自变量的后 i 个分量为 $L_n, L_{n-1}, \cdots, L_{n-i+1}$。而 $L_n, L_{n-1}, \cdots, L_{n-i+1}$ 的自变量又分别为：

$$
\begin{array}{ll}
L_n: & (\xi_1, \cdots, \xi_{n-i}, L_{n-i+1}, \cdots, L_{n-1}); \\
L_{n-1}: & (\xi_1, \cdots, \xi_{n-i}, L_{n-i+1}, \cdots, L_{n-2}); \\
\vdots & \vdots \\
L_{n-i+2}: & (\xi_1, \cdots, \xi_{n-i}, L_{n-i+1}) \\
L_{n-i+1}: & (\xi_1, \cdots, \xi_{n-i}).
\end{array}
$$

由此可见，$L_n, L_{n-1}, \cdots, L_{n-i+1}$ 在点 $x_{n-i}^{(k)}$ 和 $x_{n-i}^{(k)}+he_j\ (j=1, \cdots, n-i)$ 的值，应自 L_{n-i+1} 出发通过逐步回代求得。回代方程取

$$
L_{s+1}=\xi_{s+1}^{(k)}-\sum_{j=1}^{s}(g_{n-s,\xi_j;k}/g_{n-s,\xi_{s+1};k})(L_j-\xi_j^{(k)})
$$
$$
-g_{n-s}/g_{n-s,\xi_{s+1};k}\quad (s=n-i, \cdots, n-1), \tag{8.5.12}
$$

（注意：在计算点 $x_{n-i}^{(k)}$ 处的值时，实际上 j 只要从 $n-i+1$ 开始）。其次，在确定了 L_{n-i} 之后，在 L_n，$L_{n-1}, \cdots, L_{n-i+1}$ 中凡出现 ξ_{n-i} 之处均用 L_{n-i} 替代。

第 n 步：这时有

$$
g_n \equiv f_n(\xi_1, L_2, \cdots, L_n), \tag{8.5.13}
$$

其中 $L_i\ (i=2, 3, \cdots, n)$ 应具有如下形式：

$$
L_i=\xi_i^{(k)}-\sum_{j=1}^{i-1}(g_{n-i+1,\xi_j;k}^{(k)}/g_{n-i+1,\xi_i;k}^{(k)})(L_j-\xi_j^{(k)})
$$
$$
-g_{n-i+1}^{(k)}/g_{n-i+1,\xi_i;k}^{(k)}. \tag{8.5.14}
$$

它们当然都是变量 ξ_1 的函数（为完整起见，定义 $L_1\equiv\xi_1$，$g_1\equiv f_1$），因此 g_n 亦为单变量 ξ_1 的函数。在点 $\xi_1^{(k)}$ 近傍，以线性函数来近似替代 g_n，并由此解出

$$
\xi_1=\xi_1^{(k)}-g_n^{(k)}/g_{n,\xi_1;k}^{(k)}, \tag{8.5.15}
$$

于是求得 $\xi_1^{(k+1)}$，即 ξ_1^* 的第 $k+1$ 个近似值。现记 $L_i\equiv\xi_i^{(k+1}$ 并利用三角方程组（8.5.14）逐步回代，即可解得 $\xi_2^{(k+1)}, \cdots, \xi_n^{(k+1)}$。

按以上步骤反复迭代，直到满足迭代收敛条件(8.3.5）或（8.3.6）和（8.3.7）为止。

至此，我们将上述方法简要总结成如下算法。

算法8.5.2（Brown 方法）

1）对 $k=0,1,\cdots,m$（m 为允许的最大迭代次数）。

1.1）对 $i=1,2,\cdots,n$（n 为未知数的个数），

1.1.1）形成第 i 个函数 g_i 的各个离散偏导数 $g_{i,\xi_j,h}$，

1.1.2）寻找其中绝对值为最大的偏导数，即

$$|g_{i,\xi_{j_0},h}| = \max_j\{|g_{i,\xi_j,h}|\},$$

1.1.3）形成方程组（8.5.14）中第 i 个方程(关于 L_i）的系数。

1.2）NEXT i。

1.3）回代以求得下一个近似解 x_{k+1}。

1.4）如满足收敛条件则转 3）。

2）NEXT k。

3）END。

牛顿法每迭代一次要计算 n^2 个偏导数以及计算 n 个分量函数值。那末，Brown 方法每迭代一次又需要计算多少次函数值呢？

据以上算法可知，$i=1$ 时，需计算 f_1 的 $n+1$ 个值；$i=2$ 时，需计算 n 个 f_2 的值，依次类推。由此可知，Brown 方法每次迭代需

$$\sum_{i=2}^{n+1} i = \frac{n^2+3n}{2}$$

次函数值计算。从这里显然可以看出"逐步代入"法的优点。当然应强调指出：这里函数值计算的节省仅指原始方程组（8.1.1）中 f_i 而言。事实上，**Brown** 方法还包括好些其它函数的计算，即线性函数 L_k 的计算。但这些计算并没有计入 $\frac{1}{2}n^2+\frac{3}{2}n$ 之中。因此，只有当函数 f_i 的计算较繁复时，**Brown** 方法相对于 N_1 算法而言，计算量才有真正的节省。

因为 **Brown** 方法每次只对一个方程 $f_i=0$ 运算，而在处理下一个方程 $f_{i+1}=0$ 时，又即刻利用已得到的信息。因此，对于方程组（8.1.1）就存在一个最优次序的问题。也就是说，方程应预先作这样的最优排列：线性方程排在最前，以后方程按接近线性的程度依次排列。能够采取这种最优次序的措施，是 **Brown** 方法的一个优点。

下面我们不加证明地引入 **Brown** 方法的局部收敛定理。

定理8.5.1 设 $F: R^n \to R^n$ 满足假设 1，则必存在一闭球 $\overline{S(x^*, \delta)}$ 和正数 ε，当 $x_0 \in \overline{S(x^*, \delta)}$ 且 $\{h_j^{(k)}\}$ 以 ε 为界时，由 **Brown** 方法生成的迭代序列 $\{x_k\}$ 收敛于 x^*。如果 F 还满足假设 2 且 $h_j^{(k)}=O(\|f_i(x_k)\|)$，则 $\{x_k\}$ 至少是平方收敛的。

第八章习题

8.1 利用牛顿法和 **Brown** 方法求下列线性方程组

i) $\begin{cases} \xi_1^2+\xi_2^2=4, \\ \xi_1^2-\xi_2^2=1; \end{cases}$ ii) $\begin{cases} \xi_1^2+\xi_2^2+\xi_3^2=1, \\ 2\xi_1^2+\xi_2^2-4\xi_3=0, \\ 3\xi_1^2-4\xi_2+\xi_3^2=0; \end{cases}$

准确到四位的近似解。初始近似解分别取 $x_0=(1.6, 1.2)^T$ 和 $x_0=(0.5, 0.5, 0.5)^T$。

8.2　设

$$f(x)=\begin{cases}0, & x=0;\\ [\xi_2(\xi_1^2+\xi_2^2)^{\frac{3}{2}}]/[(\xi_1^2+\xi_2^2)^2+\xi_2^2], & x\neq 0;\end{cases}$$

证明 $f(x)$ 在 $x=0$ 处偏导数存在，但不存在 F- 导数。

8.3　设 $f(x)=\xi_1^2+\xi_1\xi_2^2$，试按定义求 $f(x)$ 的一阶和二阶 Fréchet 导数。

8.4　试证：使(8.2.14)成立的矩阵 A 至多只有一个。

8.5　设 F：$R^n\to R^n$ 在开凸集 D 上 F- 可导，试证 $F'(x)$ 在 D 内连续的充要条件是所有的偏导数 $\frac{\partial f_i}{\partial \xi_j}$ $(i,j=1,2,\cdots,n)$ 在 D 内连续。

8.6　设 $\{x_k\}\subset R^n$ 是由迭代算法 M 所生成，它收敛于(8.1.1)的解 x^*。若对某向量范数 $\|\cdot\|$ 成立等式

$$\lim_{k\to\infty}\|x_k-x^*\|^{1/k}=0,$$

则称 $\{x_k\}R$- 超线性收敛于 x^*。试证：如果 $\{x_k\}$ 商超线性收敛于 x^*，则 $\{x_k\}$ 必 R- 超线性收敛，反之不必成立。

8.7　设 $A(x)\in R^{n\times n}$ 在点 $x^*\in D$ 处连续且 $A(x^*)$ 为非异，则存在一闭球 $\overline{S(x^*,\ \delta)}\subset D$，在 $\overline{S(x^*,\ \delta)}$ 上 $A(x)^{-1}$ 在 x^* 处连续。

8.8　设 F：$D\subset R^n\to R^n$ 在 x^* 处 F- 可微，定义在 D 上的矩阵 $A(x)\in R^{n\times n}$ 在 x^* 处连续且 $A(x^*)$ 非异，则存在闭球 $\overline{S(x^*,\ \delta)}\subset D$，使向量函数

$$G(x)=x-[A(x)]^{-1}F(x)$$

在其上有定义且 $G(x)$ 在 x^* 处 F- 可微，F- 导数为

$$G'(x^*)=I-(A(x^*))^{-1}F'(x^*).$$

8.9　设 F：$R^n\to R^n$ 满足假设 1 和假设 2，试证由(8.4.26)确定的迭代序列 $\{x_k\}$ 收敛阶至少为 k_0+1。

8.10　设非线性方程组为

$$f_1(\xi_1,\xi_2)\equiv\xi_1^2+10\xi_1\xi_2+4\xi_2^2+0.7401006=0,$$

$$f_2(\xi_1,\xi_2)\equiv\xi_1^2-3\xi_1\xi_2+2\xi_2^2-1.0201228=0;$$

试用带收敛因子的牛顿法求方程组的近似解。初始近似解取 $x_0=(0.1,-0.1)^T$，$\varepsilon_1=\varepsilon_2=10^{-6}$ 或自定。要求最后打印方程组近似解，函数 f_1、

f_2 的值以及迭代次数。

8.11　设 $F: R^n \to R^n$ 满足假设 1，试证存在一闭球 $\overline{S(x^*, \delta)}$ 以及正数 β 使

$$\|F(x)\| \geqslant \beta \|x - x^*\|, \qquad \forall x \in \overline{S(x^*, \delta)}$$

成立。

8.12　编制一个离散 Brown 算法的程序，算法中

$h_j^{(k)}(j = 1, \cdots, n)$ 按如下规定选取：

$$h_j^{(k)} = \max\{\alpha_{ij}^{(k)}, \ 5 \times 10^{-\beta+2}\},$$

$$\alpha_{ij}^{(k)} = \min\{\max(|f_1^{(k)}|, |g_2^{(k)}|, \cdots, g_i^{(k)}|), \ 0.001 \times |\xi_j^{(k)}|\},$$

其中 β 为机器的有效数位。

8.13　利用上题程序求以下非线性方程组的近似解。

i）$\begin{cases} f_1(\xi_1, \xi_2) \equiv 2\xi_1^2 - \xi_2^2 - 1 = 0, \\ f_2(\xi_1, \xi_2) \equiv \xi_1^2 - \xi_2 - 4 = 0, \end{cases} \quad x_0 = (1.2, \ 1.7)^T$

ii）$\begin{cases} f_1(\xi_1, \xi_2) \equiv \cos(0.4\xi_2 + \xi_1^2) + \xi_1^2 + \xi_2^2 - 1.6 = 0, \\ f_1(\xi_1, \xi_2) \equiv 1.5\xi_1^2 - \dfrac{\xi_2^2}{0.36} - 1 = 0. \end{cases} \quad x_0 = (1.04, \ 0.47)^T,$

参 考 文 献

[1] Dennis, J.E & Moré, J.J. (1977). "Quasi-Newton Methdds, Motivation and Theory" SIAM Review. vol. 19, No. 1 January.

[2] Ortega, J. M. & W. C. Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several variables*, Academic Press, New York.

[3] Byrne, G, D. & Hall, C. A. (1972), *Numerical Solution of Systems of Nonlinear Algebric Equations Academic press New Youk.*

[4] 王德人　非线性方程组解法与最优化方法。

[5] Березин, И. С. И Жидков, Н. П. (1966), *Методы Бычислений*.

[6] Wait, R. (1979), *The Numerical Solution of Algebraic Equations.*

第九章　非线性方程组的拟牛顿迭代解法

§1　拟牛顿方法的基本思想

在第八章§4我们曾指出：牛顿法（离散牛顿法亦类似）的缺点之一是，每迭代一次需计算n^2+n个纯量函数以及$O(n^3)$次算术运算．现在随着实际问题维数的提高，迭代次数的增多，如何减少每次迭代所需的工作量就显得十分必要了．

1959、1965年，**Davidon**和**Broyden**针对无约束极小化问题以及非线性方程组（先后）提出了一类新的方法．1963年，**Fletcher**和**Powell**对**Davidan**的工作做了重要的修改和澄清，从而将这方面的研究工作引向了深入．这类新方法的基本思想是，用矩阵B_k近似替代$F'(x_k)$，而B_{k+1}可在B_k的基础上用一个低秩矩阵来校正，所需的运算量为$O(n)^2$次算术运算．显然，这样的方案计算量大大降低了．

当然计算量降低也是要付出代价的，偿还的代价就是将平方收敛速度降为超线性收敛速度．

这类新方法名称繁多，有拟牛顿法、变尺度方法、修正方法以及校正方法等等．下面我们针对非线性方程组来引入这类方法．

设$F: R^n\to R^n$满足假设1，$x_k\in D$而$s_k\neq 0$，$x_k+s_k=x_{k+1}\in D$，则据引理8.4.4知：对给定的$\varepsilon>0$，总存在$\delta>0$，当$\|x_{k+1}-x_k\|<\delta$时使

$$\|F(x_k)-F(x_{k+1})-F'(x_{k+1})(x_k-x_{k+1})\|\leqslant\varepsilon\|x_k-x_{k+1}\| \tag{9.1.1}$$

成立。因此有

$$F(x_k)\approx F(x_{k+1})+F'(x_{k+1})(x_k-x_{k+1}), \tag{9.1.2}$$

其近似程度随$\|x_k-x_{k+1}\|$的减小而增大。如果要求B_{k+1}来近似替代$F'(x_{k+1})$的话，那末很自然地要求B_{k+1}满足方程

$$F(x_k)=F(x_{k+1})+B_{k+1}(x_k-x_{k+1}) \tag{9.1.3}$$

或

$$B_{k+1}s_k=F(x_{k+1})-F(x_k). \tag{9.1.4}$$

如果$F(x)$为单变量函数，那末(9.1.4)即为

$$B_{k+1}=\frac{F(x_{k+1})-F(x_k)}{x_{k+1}-x_k}.$$

上式右端表示点x_{k+1}处$F(x)$的差商，所以(9.1.4)常称为**广义差商条件**。因为B_{k+1}可以看作$F'(x_{k+1})$的近似矩阵，所以(9.1.4)又称为拟牛顿方程，它是发展拟牛顿法的一个中心。

前面已经指出：除要求B_{k+1}满足拟牛顿方程之外，我们还希望B_{k+1}能表示为$B_k+\Delta B_k$，这里ΔB_k为一低秩矩阵。这样就得到了一类迭代算法：

$$\begin{cases} x_{k+1}=x_k-B_k^{-1}F(x_k), \\ B_{k+1}(x_{k+1}-x_k)=F(x_{k+1})-F(x_k), \\ B_{k+1}=B_k+\Delta B_k,\ \operatorname{rank}\Delta B_k=m\geqslant 1. \end{cases} \quad (k=0,1,\cdots), \tag{9.1.5}$$

如果对所有的k，$B_k^{-1}=H_k$存在，那末(9.1.5)的过程是有意义的，利用Sherman-Morrison-Woodbury公式(习题9.2)可得到ΔH_k的一个显表达式。具体地说，因为$\operatorname{rank}(\Delta B_k)=m$，而任何秩为$m$的$n$阶矩阵都可分解为

$$\Delta B_k=U_kV_k^T,\ U_k,V_k\in\mathbf{R}^{n\times m},$$
$$\operatorname{rank}U_k=\operatorname{rank}V_k=m.$$

如果$(I+V_k^TB_k^{-1}U_k)$为非异，则有

$$(B_k+U_kV_k^T)^{-1}=B_k^{-1}-B_k^{-1}U_k(I+V_k^TB_k^{-1}U_k)^{-1}V_k^TB_k^{-1},$$

于是

$$\Delta H_k=Y_kW_k^T, \tag{9.1.6}$$

其中

$$Y_k=-B_k^{-1}U_k(I+V_k^TB_k^{-1}U_k)^{-1}, \tag{9.1.7}$$

$$W_k=(B_k^{-1})^TV_k. \tag{9.1.8}$$

显然Y_k、$W_k\in\boldsymbol{R}^{n\times m}$且它们的秩均为$m$，故$\Delta H_k$仍为一个秩$m$矩阵。这样(9.1.5)又可改写为

$$\begin{cases}\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-H_k\boldsymbol{F}(\boldsymbol{x}_k),\\ H_{k+1}[\boldsymbol{F}(\boldsymbol{x}_{k+1})-\boldsymbol{F}(\boldsymbol{x}_k)]=\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\\ H_{k+1}=H_k+\Delta H_k,\ \operatorname{rank}\Delta H_k=m.\end{cases}\quad(k=0,1,\cdots) \tag{9.1.9}$$

以下很自然地要讨论ΔB_k、ΔH_k的具体计算，下节要介绍的Broyden方法，就是一种秩1校正公式。

§2 Broyden方法

为书写方便起见，以下暂将矩阵和向量的下标删去，并采用如下符号：

$$\bar{\boldsymbol{x}}=\boldsymbol{x}_{k+1},\quad \boldsymbol{x}=\boldsymbol{x}_k,\quad \boldsymbol{s}=\bar{\boldsymbol{x}}-\boldsymbol{x},\quad \boldsymbol{y}=\boldsymbol{F}(\bar{\boldsymbol{x}})-\boldsymbol{F}(\boldsymbol{x}),$$

$$\bar{B}=B_{k+1},\ \bar{H}=H_{k+1},\ B=B_k,\quad H=H_k,$$

$$\Delta B=\bar{B}-B,\ \Delta H=\bar{H}-H$$

$$\boldsymbol{u}=\boldsymbol{y}-B\boldsymbol{s},\ \boldsymbol{v}=-H\boldsymbol{u}=\boldsymbol{s}-H\boldsymbol{y}.$$

现在考虑校正矩阵ΔB为秩1阵的情形。因为任何秩1阵均可表为二向量的外积，故设

$$\Delta B=\boldsymbol{b}\boldsymbol{c}^T\qquad \boldsymbol{b}、\boldsymbol{c}\in\boldsymbol{R}^n \tag{9.2.1}$$

或

$$\bar{B} = B + \boldsymbol{b}\boldsymbol{c}^T \tag{9.2.2}$$

代入拟牛顿方程即得

$$(B + \boldsymbol{b}\boldsymbol{c}^T)\boldsymbol{s} = \boldsymbol{y},$$

故得

$$\boldsymbol{u} = \langle \boldsymbol{c}, \boldsymbol{s}\rangle \boldsymbol{b}.$$

于是秩1校正公式应在下面范围中选取

$$\Delta B = \sigma \boldsymbol{u}\boldsymbol{c}^T,$$

而

$$\sigma = \begin{cases} 1/\langle \boldsymbol{c}, \boldsymbol{s}\rangle, & \boldsymbol{s} \neq 0; \\ 0, & \boldsymbol{s} = 0, \end{cases}$$

或

$$\bar{B} = B + \frac{(\boldsymbol{y} - B\boldsymbol{s})\boldsymbol{c}^T}{\langle \boldsymbol{c}, \boldsymbol{s}\rangle} \quad (\boldsymbol{s} \neq 0). \tag{9.2.3}$$

因为我们把$\bar{B}$看作$F'(\bar{\boldsymbol{x}})$的一个近似，所以确定$\bar{B}$的条件应是$\boldsymbol{s}$方向上的导数即拟牛顿方程，而在与$\boldsymbol{s}$正交的方向上我们未曾提出过什么要求。所以，在与$\boldsymbol{s}$正交的所有方向$\boldsymbol{z}$上要求成立

$$\bar{B}\boldsymbol{z} = B\boldsymbol{z}, \quad \forall \boldsymbol{z} \in Z \equiv \{\boldsymbol{z} \mid \boldsymbol{z} \neq 0, \ \langle \boldsymbol{z}, \boldsymbol{s}\rangle = 0\} \tag{9.2.4}$$

这种想法是合乎逻辑的。将(9.2.3)和(9.2.4)结合起来，即可得到唯一的校正公式

$$\Delta B = \sigma \boldsymbol{u}\boldsymbol{s}^T,$$

$$\delta = \begin{cases} 1/\langle \boldsymbol{s}, \boldsymbol{s}\rangle, & \boldsymbol{s} \neq 0 \\ 0, & \boldsymbol{s} = 0 \end{cases}$$

或

$$\bar{B} = B + \frac{(\boldsymbol{y} - B\boldsymbol{s})\boldsymbol{s}^T}{\langle \boldsymbol{s}, \boldsymbol{s}\rangle}. \quad (\boldsymbol{s} \neq 0). \tag{9.2.5}$$

这样，我们就得到了 **Broyden** 秩 1 迭代算法。

算法9.2.1 设 $\boldsymbol{F}: \boldsymbol{R}^n \to \boldsymbol{R}^n$，$\boldsymbol{x}_0$为$\boldsymbol{x}^*$的初始近似解，$B_0$为初始迭代矩阵（通常 B_0 就取作 $\boldsymbol{F}'(\boldsymbol{x}_0)$），它们都符合定理9.3.1的假定。

1) 计算 $\boldsymbol{F}(\boldsymbol{x}_0)$。

2) 对$k=0, 1, \cdots, m$，

2.1) 求解线性方程组，

$$B_k \Delta \boldsymbol{x}_k = -\boldsymbol{F}(\boldsymbol{x}_k)$$

并置$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \Delta \boldsymbol{x}_k \qquad (\Delta \boldsymbol{x}_k \equiv \boldsymbol{s}_k)$，

2.2) 计算 $\boldsymbol{F}(\boldsymbol{x}_{k+1})$，如果 $\boldsymbol{x}_{k+1}$满足精度要求，那末转4)，

2.3) 计算$\boldsymbol{y}_k = \boldsymbol{F}(\boldsymbol{x}_{k+1}) - \boldsymbol{F}(\boldsymbol{x}_k)$，

2.4) 计算B_{k+1}

$$B_{k+1} = B_k + \frac{(\boldsymbol{y}_k - B_k \boldsymbol{s}_k)\boldsymbol{s}_k^T}{(\boldsymbol{s}_k, \ \boldsymbol{s}_k)}.$$

3) NEXT k。

4) 置$\boldsymbol{x}^* = \boldsymbol{x}_{k+1}$。

5) END。

Broyden 校正公式具有一个十分重要的性质，即在所有满足拟牛顿方程的矩阵集合中，$\bar{B}$是在Frobenius范数意义下最接近B的矩阵.实际上，这一性质等价于确定$\bar{B}$的条件(9.2.4)。

定理9.2.1 设给定$B \in \boldsymbol{R}^{n \times n}$，$\boldsymbol{y} \in \boldsymbol{R}^n$以及非零向量$\boldsymbol{s} \in \boldsymbol{R}^n$，$\bar{B}$为由(9.2.5)确定的矩阵，则$\bar{B}$为以下问题的唯一解。

$$\min\{\|\hat{B} - B\|_F \mid \hat{B}\boldsymbol{s} = \boldsymbol{y}\}. \tag{9.2.6}$$

证 利用(9.2.5)并考虑到$\boldsymbol{y} = \hat{B}\boldsymbol{s}$及习题9.1有

$$\|\bar{B} - B\|_F = \left\|(\hat{B} - B)\frac{\boldsymbol{s}\boldsymbol{s}^T}{(\boldsymbol{s}, \boldsymbol{s})}\right\|_F$$

$$\leqslant \|\hat{B}-B\|_F \left\| \frac{ss^T}{\langle s, s\rangle} \right\|_2$$

$$= \|\hat{B}-B\|_F,$$

所以 $\bar{B}$ 确为问题(9.2.6)的解。至于唯一性可以这样说明：设 $f(A)=\|A-B\|_F$，不难验证函数 $f(A)$ 在集合

$$N=\{A \mid As=y\} \tag{9.2.7}$$

上是一个严格凸函数*。另一方面，集合 N 显然是 $R^{n\times n}$ 中的一个凸集。因此，$\bar{B}$ 为凸函数 $f(A)$ 在凸集 N 上的唯一解(习题9.4)。

对校正公式(9.2.5)利用Sherman-Morrison公式可得

$$\bar{H}=H-\frac{Hbs^TH}{1+s^THb},$$

其中 $b=\sigma u$。因为 $\sigma Hy=\sigma s+Hb$，故 $1+s^THb=\sigma s^THy$。将它们代入上式即可得

$$\bar{H}=H-\frac{Hus^TH}{s^THy}=H+\frac{(s-Hy)s^TH}{s^THy}. \tag{9.2.9}$$

由此可见，如果 $B^{-1}=H$ 存在，那末，$\bar{B}$ 为非异的充要条件为 $s^THy\neq 0$。

算法9.2.2 设 F: $R^n\to R^n$，x_0 为 x^* 的初始近似解，H_0 为初始迭代矩阵(通常 H_0 取作 $F'(x_0)^{-1}$，它们都符合定理9.3.1的假设。

1) 计算 $F(x_0)$ 和 $F'(x_0)$。

2) 对 $k=0,1,\cdots,m$，

---

*函数 f: $R^n\to R$ 在凸集 $D\subset R^n$ 上是凸的，意指对任意 $x', x''\in D$ 且 $x'\neq x''$ 以及任意的 $\lambda\in(0,1)$ 成立

$$f[\lambda x'+(1-\lambda)x'']\leqslant \lambda f(x')+(1-\lambda)f(x'') \tag{9.2.8}$$

如果"$\leqslant$"改为"$<$"，那末函数 f 在 D 上是严格凸的。

2.1) 计算$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-H_k\boldsymbol{F}(\boldsymbol{x}_k)\quad(-H_k\boldsymbol{F}(\boldsymbol{x}_k)=\boldsymbol{s}_k)$，

2.2) 计算 $\boldsymbol{F}(\boldsymbol{x}_{k+1})$，如果 $\boldsymbol{x}_{k+1}$满足精度要求，那末转4)，

2.3) 计算$\boldsymbol{y}_k=\boldsymbol{F}(\boldsymbol{x}_{k+1})-\boldsymbol{F}(\boldsymbol{x}_k)$，

2.4) 计算H_{k+1}

$$H_{k+1}=H_k+\frac{(\boldsymbol{s}_k-H_k\boldsymbol{y}_k)\boldsymbol{s}_k^TH_k}{\boldsymbol{s}_k^TH_k\boldsymbol{y}_k}\ .$$

3) *NEXT* k。

4) 置 $\boldsymbol{x}^*=\boldsymbol{x}_{k+1}$。

5) END。

比较算法9.2.1和算法9.2.2，当给定$\boldsymbol{x}_0$和B_0后，算法9.2.1每迭代一次，要计算n个纯量函数，然后求解线性方程组，因此每迭代一次还需$O(n^3)$次算术运算。算法9.2.2每迭一次亦需计算n个纯量函数。另外，形成校正公式(9.2.9)需$O(n^2)$次算术运算。由此可见，采用算法9.2.2较好。可是，利用校正公式(9.2.5)也有可能迭代一次只需$O(n^2)$次算术运算。1972年，Gill和Murray提出了一种计算方案：设已知$B_k=Q_kR_k$(这里Q_k为正交阵，R_k为上三角阵)，则形成B_{k+1}的QR分解也只要$O(n^2)$次算术运算。当然，如果$B_k=Q_kR_k$，那末，求解线性方程组$B_k\boldsymbol{s}_k=-\boldsymbol{F}(\boldsymbol{x}_k)$就只需$O(n^2)$次算术运算。这个计算方案要比算法9.2.2好，这是因为校正公式(9.2.5)中毋需执行矩阵与向量的乘法，因为$B_k\boldsymbol{s}_k$即为$-\boldsymbol{F}(\boldsymbol{x}_k)$。另外，理论分析也表明，校正公式(9.2.5)更为稳定。

例9.2.1 利用Broyden秩1方法求方程组

$$f_1(\xi_1,\xi_2)\equiv\xi_1^2-\xi_2-1=0,$$

$$f_2(\xi_1,\xi_2)\equiv(\xi_1-2)^2+(\xi_2-0.5)^2-1=0$$

的近似解。初始近似解取$\boldsymbol{x}_0=(0,0)^T$。

解 采用算法9.2.2计算。首先计算 $F(x_0)=(-1,3.25)^T$。因为

$$F'(x_0)=\begin{pmatrix}2\xi_1 & -1\\ 2\xi_1-4 & 2\xi_2-1\end{pmatrix},$$

故

$$F'(x_0)=\begin{pmatrix}0 & -1\\ -4 & -1\end{pmatrix},\quad H_0=F'(x_0)^{-1}=\begin{pmatrix}0.25 & -0.25\\ -1 & 0\end{pmatrix}.$$

据牛顿法求得，$x_1=(1.0625,\ -1)^T$；$F(x_1)=(1.12890625,\ 2.12890625)^T$；按语句2.3)和2.4)算得

$$s_1=\begin{pmatrix}1.0652\\ -1\end{pmatrix},\quad y_1=\begin{pmatrix}2.12890625\\ -1.12109375\end{pmatrix}$$

以及

$$H_1=\begin{pmatrix}0.3557441\cdots & -0.2721932\cdots\\ -0.5224991\cdots & -0.1002162\cdots\end{pmatrix}.$$

按语句2.1)算得

$$x_2=\begin{pmatrix}1.0625\\ -1\end{pmatrix}-\begin{pmatrix}0.3557441\cdots & -0.2721932\cdots\\ -0.5224991\cdots & -0.1002162\cdots\end{pmatrix}$$

$$\begin{pmatrix}1.12890625\\ 2.12890625\end{pmatrix}=\begin{pmatrix}1.240372062\cdots\\ -0.1967974670\cdots\end{pmatrix}.$$

经过11次迭代可得到具有12位有效数字的近似解：

$$x_{11}=(1.54634288332,\ 1.39117631279)^T.$$

如果用牛顿法，那末只要经七次迭代即能得到这一结果。因次，就一般而言，Broyden 方法的收敛速度要较牛顿方法收敛速度慢。当然，对两个联立方程的情形，Broyden 方法的优点显得不突出，甚至不如用牛顿法方便。只有当 n 较大时，在计算上才能显出它的优点。得到这一好处的代价就是收敛速度较牛顿法要稍慢一点。下一节我们就来分析 Broyden方法的收敛性。

§3 Broyden方法的收敛性分析

首先应该强调指出：以下采用的证明技巧也适用于其它拟牛顿方法的收敛性分析。现假定F满足假设1和假设2，则当初始近似解x_0和初始矩阵B_0分别充分接近x^*和$F'(x^*)$时，算法9.2.1生成的迭代序列$\{x_k\}$超线性收敛于方程组的解x^*。为了证明这一结论，首先证明以下几个引理。

引理9.3.1 设F：$R^n \to R^n$满足假设1和假设2，$x \in D$而$\bar{x}$由算法9.2.1生成且设$\bar{x} \in D$，则

$$\|\bar{B} - F'(x^*)\|_p \leqslant \|B - F'(x^*)\|_p + \kappa\sigma(x, \bar{x}), \ (p = 2, F) \tag{9.3.1}$$

其中

$$\sigma(x, \bar{x}) = \max\{\|\bar{x} - x^*\|, \|x - x^*\|\}. \tag{9.3.2}$$

证 由(9.2.5)有

$$\bar{B} - F'(x^*) = (B - F'(x^*))\left(I - \frac{s s^T}{\langle s, s \rangle}\right) + \frac{(y - F'(x^*)s)s^T}{\langle s, s \rangle}, \tag{9.3.3}$$

上式两端取范数，据习题9.1～9.3即有

$$\|\bar{B} - F'(x^*)\|_p \leqslant \|B - F'(x)\|_p + \frac{\|y - F'(x^*)s\|}{\|s\|}.$$

据引理8.4.4有

$$\|y - F'(x^*)s\| \leqslant \kappa\sigma(x, \bar{x})\|s\|$$

成立。至此(9.3.1)得证。

引理9.3.2 设F：$R^n \to R^n$满足假设1和假设2，又设$\{B_k\}$为一非异矩阵序列。如果对某$x_0 \in D$迭代序列

$$x_{k+1}=x_k-B_k^{-1}F(x_k) \quad (k=0,1,\cdots)$$

恒在D中且$x_k\neq x^*$($\forall k\leqslant 0$)，又设该迭代序列收敛于x^*，那末当且仅当

$$\lim_{k\to\infty}\frac{\|[B_k-F'(x^*)](x_{k+1}-x_k)\|}{\|x_{k+1}-x_k\|}=0 \qquad (9.3.4)$$

时，$\{x_k\}$超线性收敛于x^*。

证 若(9.3.4)成立，则因

$$\begin{aligned}[B_k-F(x^*)](x_{k+1}-x_k)&=-F(x_k)-F'(x^*)(x_{k+1}-x_k)\\&=[F(x_{k+1})-F(x_k)-F'(x^*)(x_{k+1}-x_k)]\\&\quad -F(x_{k+1}),\end{aligned} \qquad (9.3.5)$$

于是

$$\lim_{k\to\infty}\frac{\|F(x_{k+1})\|}{\|x_{k+1}-x_k\|}\leqslant K\lim_{k\to\infty}\sigma(x_k,x_{k+1})$$

$$+\lim_{k\to\infty}\frac{\|[B_k-F'(x^*)](x-x_k)\|}{\|x_{k+1}-x_k\|},$$

故

$$\lim_{k\to\infty}\frac{\|F(x_{k+1})\|}{\|x_{k+1}-x_k\|}=0。 \qquad (9.3.6)$$

又因$F'(x^*)$为非异，故存在一正数β，当k充分大以后有

$$\|F(x_{k+1})\|=\|F(x_{k+1})-F(x^*)\|\geqslant\beta\|x_{k+1}-x^*\|$$

(习题8.11)，据假定，对所有$k\geqslant 0$，$x_k\neq x^*$且x_k恒在D中，故当k充分大后有

$$\frac{\|F(x_{k+1})\|}{\|x_{k+1}-x_k\|}\geqslant\frac{\beta\|x_{k+1}-x^*\|}{\|x_{k+1}-x^*\|+\|x_k-x^*\|}$$

$$=\beta\frac{q_k}{1+q_k}。$$

其中 $q_k=\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|/\|\boldsymbol{x}_k-\boldsymbol{x}^*\|$。据(9.3.6)知 $q_k/(1+q_k)$ 必收敛于零，因此 q_k 亦必收敛于零。据定义 8.3.2知$\{\boldsymbol{x}_k\}$超线性收敛于$\boldsymbol{x}^*$。

反之，若$\{\boldsymbol{x}_k\}$超线性收敛于$\boldsymbol{x}^*$且$\boldsymbol{F}(\boldsymbol{x}^*)=0$，则因

$$\frac{\|\boldsymbol{F}(\boldsymbol{x}_{k+1})\|}{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|}=\frac{\|\boldsymbol{F}(\boldsymbol{x}_{k+1})-\boldsymbol{F}(\boldsymbol{x}^*)\|}{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|}\cdot\frac{\|\boldsymbol{x}_k-\boldsymbol{x}^*\|}{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|},$$

于是利用引理8.4.3及(8.3.3)、(8.3.4)式即知

$$\lim_{k\to\infty}\frac{\|\boldsymbol{F}(\boldsymbol{x}_{k+1})\|}{\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|}=0,$$

再利用(9.3.5)式即可推知(9.3.4)式。

定理9.3.1 设$\boldsymbol{F}$：$\boldsymbol{R}^n\to\boldsymbol{R}^n$ 满足假设1和假设2，则 Broyden 算法 9.2.1 局部、超线性收敛。换言之，存在正数 ε 和 δ，当 $\|B_0-\boldsymbol{F}'(\boldsymbol{x}^*)\|\leqslant\delta$、$\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\leqslant\varepsilon$ 时，$\{\boldsymbol{x}_k\}$ 超线性收敛于 $\boldsymbol{x}^*$。

证 设$\alpha=\|\boldsymbol{F}'(\boldsymbol{x}^*)^{-1}\|$，取 $\delta\leqslant\frac{1}{6\alpha}$，$\varepsilon\leqslant\frac{\delta}{2\kappa}$ 且保证$\overline{S(\boldsymbol{x}^*,\varepsilon)}\subset D$。现取 $\boldsymbol{x}_0$、B_0满足定理要求，于是据 Banach 引理 B_0 必为非异且

$$\|B_0^{-1}\|\leqslant\alpha(1-\alpha\delta)^{-1}=\gamma。$$

又ε的选取方法保证了 $\boldsymbol{F}(\boldsymbol{x}_0)$存在，因此 $\boldsymbol{x}_1$ 是完全确定的。据引理8.4.4及定理假设应有

$$\begin{aligned}\|\boldsymbol{x}_1-\boldsymbol{x}^*\|&=\|\boldsymbol{x}_0-\boldsymbol{x}^*-B_0^{-1}(\boldsymbol{F}(\boldsymbol{x}_0)-\boldsymbol{F}(\boldsymbol{x}^*))\|\\&\leqslant\|-B_0^{-1}\|\|\boldsymbol{F}(\boldsymbol{x}_0)-\boldsymbol{F}(\boldsymbol{x}^*)-B_0(\boldsymbol{x}_0-\boldsymbol{x}^*)\|\\&\leqslant\|B_0^{-1}\|[\|\boldsymbol{F}(\boldsymbol{x}_0)-\boldsymbol{F}(\boldsymbol{x}^*)-\boldsymbol{F}'(\boldsymbol{x}^*)(\boldsymbol{x}_0-\boldsymbol{x}^*)\|\\&\quad+\|\boldsymbol{F}'(\boldsymbol{x}^*)-B_0\|\|\boldsymbol{x}_0-\boldsymbol{x}^*\|]\\&\leqslant\gamma(\frac{1}{2}\kappa\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2+\delta\|\boldsymbol{x}_0-\boldsymbol{x}^*\|)\end{aligned}$$

$$\leqslant \gamma\left(\frac{\delta}{4}+\delta\right)\|\boldsymbol{x}_0-\boldsymbol{x}^*\|$$

$$<\frac{1}{2}\|\boldsymbol{x}_0-\boldsymbol{x}^*\|.$$

由此可见$\boldsymbol{x}_1\in\overline{S(\boldsymbol{x}^*,\varepsilon)}$，故$\boldsymbol{F}(\boldsymbol{x}_1)$和$B_1$存在。再利用引理9.3.1即可推得

$$\begin{aligned}\|\boldsymbol{B}_1-\boldsymbol{F}'(\boldsymbol{x}^*)\| &\leqslant \|B_0-\boldsymbol{F}'(\boldsymbol{x}^*)\|+\kappa\sigma(\boldsymbol{x}_0,\boldsymbol{x}_1)\\ &\leqslant \delta+\kappa\|\boldsymbol{x}_0-\boldsymbol{x}^*\|\\ &\leqslant \delta+\frac{1}{2}\delta=\frac{3}{2}\delta<2\delta.\end{aligned}$$

于是据 Banach 引理，B_1 为非异且

$$\|B_1^{-1}\|\leqslant \alpha(1-2\alpha\delta)^{-1},$$

因此$\boldsymbol{x}_2$是可以确定的。现作归纳法假定：设$\boldsymbol{x}_1$，…，$\boldsymbol{x}_n$和B_1^{-1}，…，B_{n-1}^{-1}均存在且当$k\leqslant n$时有

$$\|\boldsymbol{x}_k-\boldsymbol{x}^*\|\leqslant\frac{1}{2}\|\boldsymbol{x}_{k-1}-\boldsymbol{x}^*\|,\quad \|B_k-\boldsymbol{F}'(\boldsymbol{x}^*)\|\leqslant\left(2-\left(\frac{1}{2}\right)^k\right)\delta.$$

于是，据Banach引理知，B_n为非异且

$$\|B_n^{-1}\|\leqslant\alpha(1-2\alpha\delta)^{-1},$$

这就保证了$\boldsymbol{x}_{n+1}$存在。另一方面

$$\begin{aligned}&\|\boldsymbol{x}_{n+1}-\boldsymbol{x}^*\|\leqslant\|B_n^{-1}\|[\|\boldsymbol{F}(\boldsymbol{x}_n)-\boldsymbol{F}(\boldsymbol{x}^*)-\boldsymbol{F}'(\boldsymbol{x}^*)(\boldsymbol{x}_n-\boldsymbol{x}^*)\|\\ &+\|\boldsymbol{F}'(\boldsymbol{x}^*)-B_n\|\|\boldsymbol{x}_n-\boldsymbol{x}^*\|]\\ &\leqslant\alpha(1-2\alpha\delta)^{-1}\left[\frac{1}{2}\kappa\|\boldsymbol{x}_n-\boldsymbol{x}^*\|^2+\left(2-\left(\frac{1}{2}\right)^n\right)\delta\|\boldsymbol{x}_n-\boldsymbol{x}^*\|\right]\\ &\leqslant\alpha(1-2\alpha\delta)^{-1}\left[\left(\frac{1}{2}\right)^n\left(\frac{\delta}{4}\right)+\left(2-\frac{1}{2}\right)^n\delta\right]\|\boldsymbol{x}_n-\boldsymbol{x}^*\|\\ &<\alpha(1-2\alpha\delta)^{-1}2\delta\|\boldsymbol{x}_n-\boldsymbol{x}^*\|\\ &<\frac{1}{2}\|\boldsymbol{x}_n-\boldsymbol{x}^*\|.\end{aligned}$$

故$\boldsymbol{x}_{n+1}\in S(\boldsymbol{x}^*,\varepsilon)$。其次

$$\|B_{n+1}-F'(x^*)\|\leqslant\|B_n-F'(x^*)\|+\kappa\sigma(x_n,x_{n+1})$$
$$\leqslant\left(2-\left(\frac{1}{2}\right)^n\right)\delta+\kappa\|x_n-x^*\|$$
$$\leqslant\left[\left(2-\left(\frac{1}{2}\right)^n\right)+\left(\frac{1}{2}\right)^{n+1}\right]\delta$$
$$=\left[2-\left(\frac{1}{2}\right)^{n+1}\right]\delta.$$

至此归纳证明完成。

综上所述，序列 $\{x_k\}$、$\{B_k^{-1}\}$ 均存在且 $\|x_k-x^*\|\leqslant\left(\frac{1}{2}\right)^k\|x_0-x^*\|$，因此 Broyden 方法至少是线性收敛的。

以下进一步证明 Broyden 方法是超线性收敛的，据引理 9.3.2知只要证明(9.3.4)式成立即可。

通过直接计算即可证明

$$\left\|E\left(I-\frac{ss^T}{(s,s)}\right)\right\|_F^2=\|E\|_F^2-\left(\frac{\|Es\|}{\|s\|}\right)^2, \tag{9.3.7}$$

其中 $E\in R^{n\times n}$。又由不等式 $(\alpha^2-\beta^2)^{1/2}\leqslant\alpha-(2\alpha)^{-1}\beta^2\ (0\leqslant\beta\leqslant\sqrt{2}\alpha)$ 即可推得

$$\left\|E\left(I-\frac{ss^T}{(s,s)}\right)\right\|_F\leqslant\|E\|_F-(2\|E\|_F)^{-1}\left(\frac{\|Es\|}{\|s\|}\right)^2. \tag{9.3.8}$$

现令 $E=B_k-F'(x^*)$ 而 $\|E\|_F=\eta_k$。对(9.3.3)利用不等式(9.3.8)和引理 8.4.4 即可推得

$$\eta_{k+1}\leqslant[1-(2\eta_k^2)^{-1}\psi_k^2]\eta_k+\kappa\sigma_k,$$

其中

$$\psi_k=\frac{\|[B_k-F'(x^*)]s_k\|}{\|s_k\|}.$$

因为$\{x_k\}$线性收敛，故$\sum_{k=0}^{\infty}\|x_k - x^*\| < +\infty$。另一方面，由引理9.3.1知，$\eta_{k+1} \leqslant \eta_k + \kappa\sigma_k$，据此不等式即可推知数列$\{\eta_k\}$有界。设$\eta$为它的某一上界，则对所有$k$均成立

$$(2\eta^2)^{-1}\psi_k^2 \leqslant \eta_k - \eta_{k+1} + \kappa\sigma_k,$$

于是

$$(2\eta^2)^{-1}\sum_{k=0}^{\infty}\psi_k^2 \leqslant \eta_0 + \kappa\sum_{k=0}^{\infty}\sigma_k$$

$$\leqslant \delta + \kappa\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k\|x_0 - x^*\| \leqslant \delta + 2\kappa\varepsilon.$$

由此可见$\{\psi_k\}$收敛于零，亦即(9.3.4)式成立。

第九章习题

9.1 设$a, b \in R^n$，试证：$\|ab^T\|_2 = \|a\|_2\|b\|_2 = \|ab^T\|_F$.

9.2 设$a, b \in R^n$且$a^Tb = 1$，试证：

i) $\|I - ab^T\|_2 = \|a\|_2\|b\|_2$；

ii) $\|I - ab^T\|_2 = 1$的充要条件为a和b共线.

9.3 设$A, B \in R^{n\times n}$，试证：

$$\|AB\|_F \leqslant \min\{\|A\|_2\|B\|_F, \|A\|_F\|B\|_2\}.$$

9.4 设给定非零向量s、$y \in R^n$，A、$B \in R^{n\times n}$，试验证：

$$f(A) = \|A - B\|_F$$

为凸集$N = \{A \mid As = y\}$上严格的凸函数.

9.5 设$A \in R^{n\times n}$为非异，$u, v \in R^n$，试证：

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TAu},$$

其中$1 + v^TAu \neq 0$ (Sherman-Morrison公式).

9.6 设$B \in R^{n\times n}$，$S \in R^{k\times k}$而$U, V \in R^{n\times k}$. 假定B、S和$V^TB^{-1}U - S^{-1}$为非异，试证：

$$(B + USV^T)^{-1} = B^{-1} - B^{-1}UTV^TB^{-1}$$

其中$S^{-1}+T^{-1}=V^TB^{-1}U$(Sherman-Morrison-Woodbury公式)。

9.7 设$F: R^n \to R^n$满足引理9.3.2的假设，如果对某$x_0 \in D$出

$$x_{k+1}=x_k-\lambda_k B_k^{-1}F(x_k) \qquad (k=0,1,\cdots)$$

生成的迭代序列$\{x_k\}$是完全确定的且$\{x_k\}\subset D$收敛于x^*，又假定(9.3.4)式成立，那末$\{x_k\}$超线性收敛于x^*，当且仅当$\{\lambda_k\}$收敛于1.

9.8 利用算法9.2.1计算例9.2.1.

9.9 试证等式(9.3.7).

参 考 文 献

1. Dennis, J.E.& More′, J.J.(1977), "*Quasi Newton Methods, Motivation and Theory*", SIAM Review. Vol.19, No.1, January.
2. Broyden, G.G. Dennis, J.E. & Moré, J.J.(1973), "*On the local and superlinear Convergence of quasi-Newton methods*", Ibid., 12, pp.223—246.
3. Dennis, J.E. & More, J. J. (1974), "*A Characterization of superlineur Convergence and its application to quasi-Newton methods*", Math Comp, 28, pp.549—560.
4. 王德人 非线性方程组解法与最优化方法。
5. Ralston, A. & Rabinowitz, P. (1978), *A First Course in Numerical Analysis*.
6. Gill, P.E. & Murray, W. (1972), "*Quasi-Newton methods for unconstrained minimization*", J.Inst. Math. Appl., 9, pp.91—108.
7. Stewart, G.w., (1973). *Introduction to Matrix Computation*. New York, Academic Press.